# Demographic inference based on Site frequency spectrum (SFS) – Part II

Vitor Sousa

**CE3C** – center for ecology, evolution and environmental changes

2018 WSPG Cesky Krumlov
26 Jan 2018

vmsousa@fc.ul.pt

UNIVERSITÄT BERN

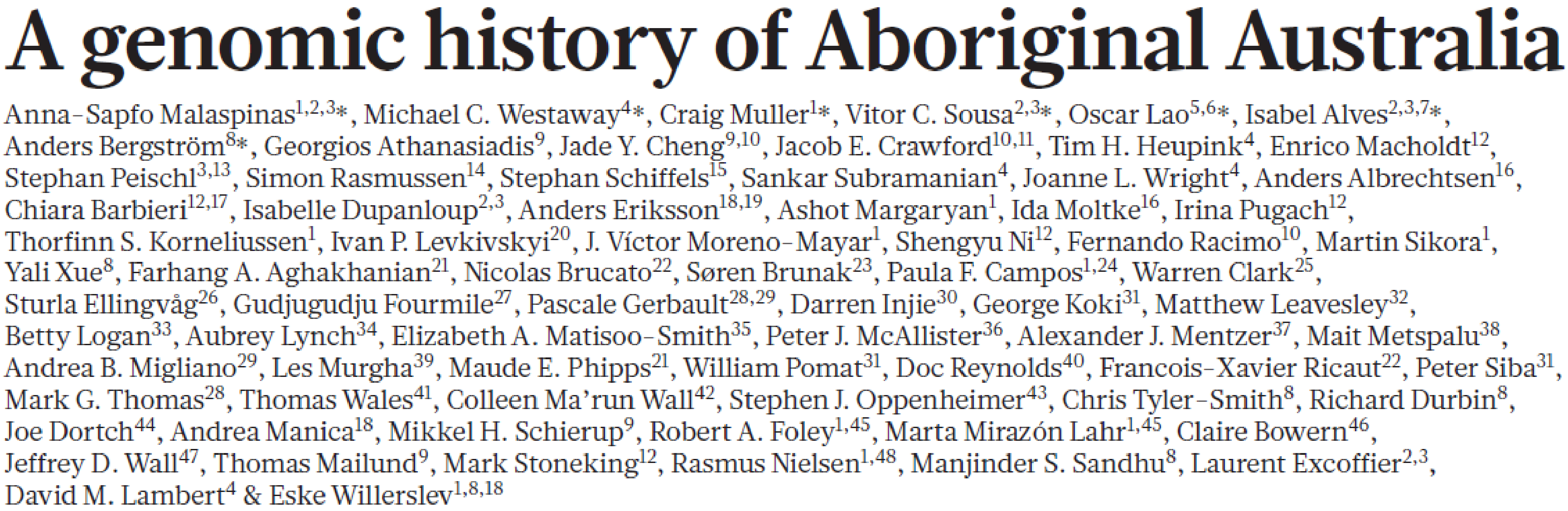
centre for ecology, evolution and environmental changes

SIB Swiss Institute of Bioinformatics

# Outline part II

Example of Applications:

- Human dispersal out of Africa (high quality whole-genome) – lessons on choice of models
- Deer mice colonization of Nebraska Sand Hills (targeted re-capture data) – lessons on effects of filtering
- Inferring divergence times and gene flow in sawflies (ddRAD-seq data) – lessons from comparing models

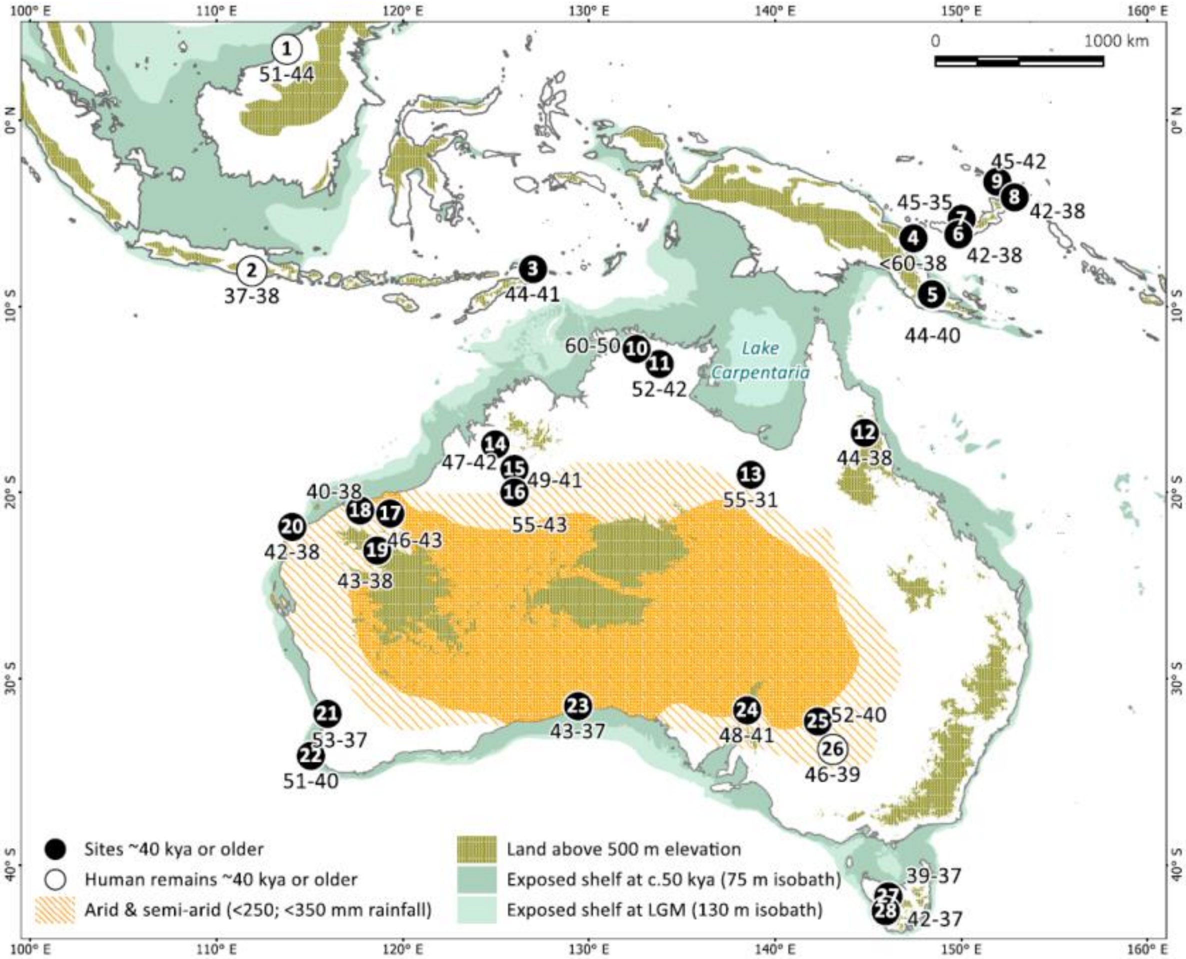


# A genomic history of Aboriginal Australia

Anna-Sapfo Malaspinas[1,2,3]*, Michael C. Westaway[4]*, Craig Muller[1]*, Vitor C. Sousa[2,3]*, Oscar Lao[5,6]*, Isabel Alves[2,3,7]*, Anders Bergström[8]*, Georgios Athanasiadis[9], Jade Y. Cheng[9,10], Jacob E. Crawford[10,11], Tim H. Heupink[4], Enrico Macholdt[12], Stephan Peischl[3,13], Simon Rasmussen[14], Stephan Schiffels[15], Sankar Subramanian[4], Joanne L. Wright[4], Anders Albrechtsen[16], Chiara Barbieri[12,17], Isabelle Dupanloup[2,3], Anders Eriksson[18,19], Ashot Margaryan[1], Ida Moltke[16], Irina Pugach[12], Thorfinn S. Korneliussen[1], Ivan P. Levkivskyi[20], J. Víctor Moreno-Mayar[1], Shengyu Ni[12], Fernando Racimo[10], Martin Sikora[1], Yali Xue[8], Farhang A. Aghakhanian[21], Nicolas Brucato[22], Søren Brunak[23], Paula F. Campos[1,24], Warren Clark[25], Sturla Ellingvåg[26], Gudjugudju Fourmile[27], Pascale Gerbault[28,29], Darren Injie[30], George Koki[31], Matthew Leavesley[32], Betty Logan[33], Aubrey Lynch[34], Elizabeth A. Matisoo-Smith[35], Peter J. McAllister[36], Alexander J. Mentzer[37], Mait Metspalu[38], Andrea B. Migliano[29], Les Murgha[39], Maude E. Phipps[21], William Pomat[31], Doc Reynolds[40], Francois-Xavier Ricaut[22], Peter Siba[31], Mark G. Thomas[28], Thomas Wales[41], Colleen Ma'run Wall[42], Stephen J. Oppenheimer[43], Chris Tyler-Smith[8], Richard Durbin[8], Joe Dortch[44], Andrea Manica[18], Mikkel H. Schierup[9], Robert A. Foley[1,45], Marta Mirazón Lahr[1,45], Claire Bowern[46], Jeffrey D. Wall[47], Thomas Mailund[9], Mark Stoneking[12], Rasmus Nielsen[1,48], Manjinder S. Sandhu[8], Laurent Excoffier[2,3], David M. Lambert[4] & Eske Willerslev[1,8,18]

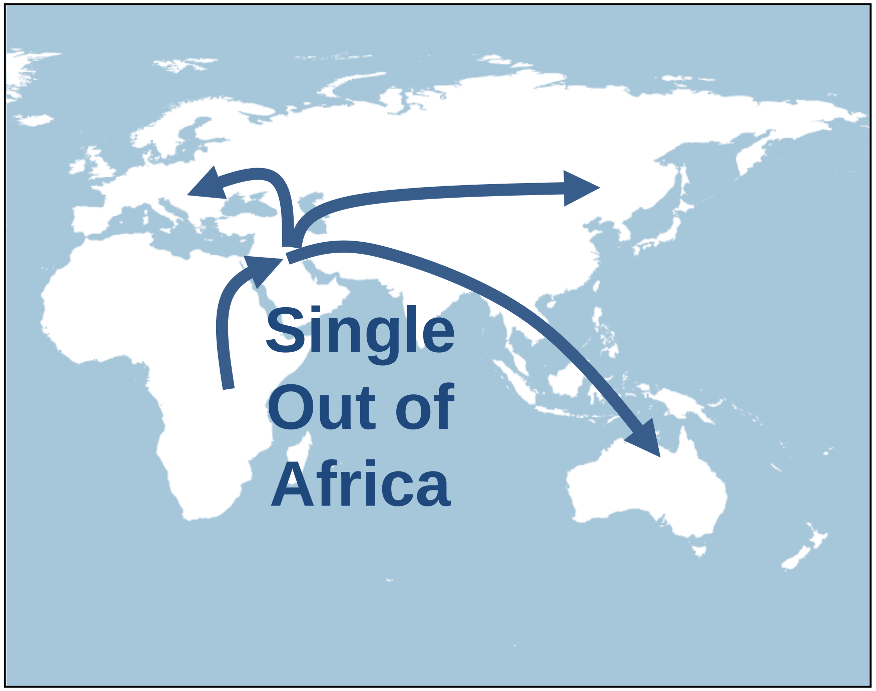
***Nature***(2016)

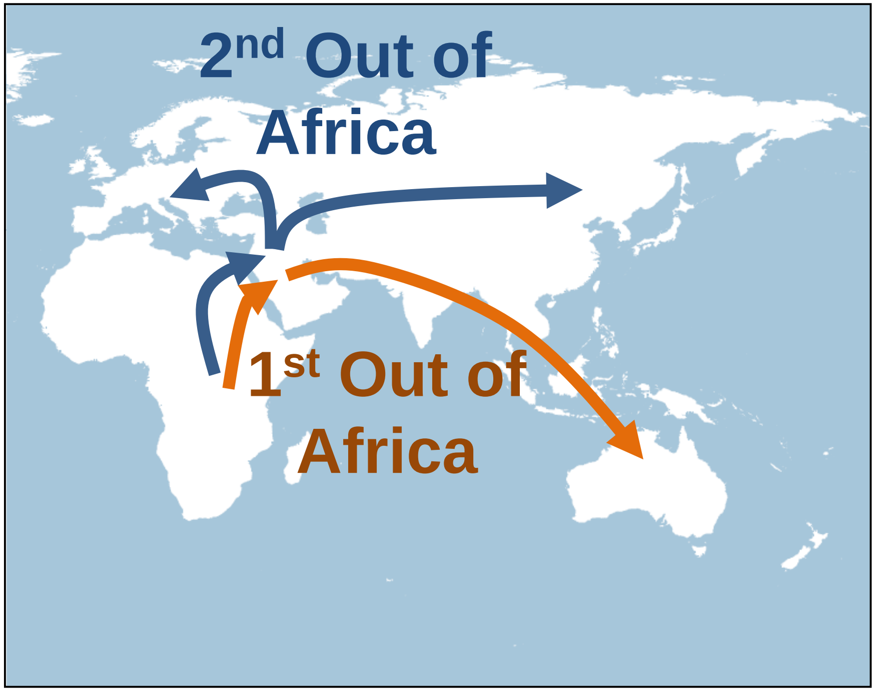
# Australia harbors some of the oldest modern human remains outside Africa



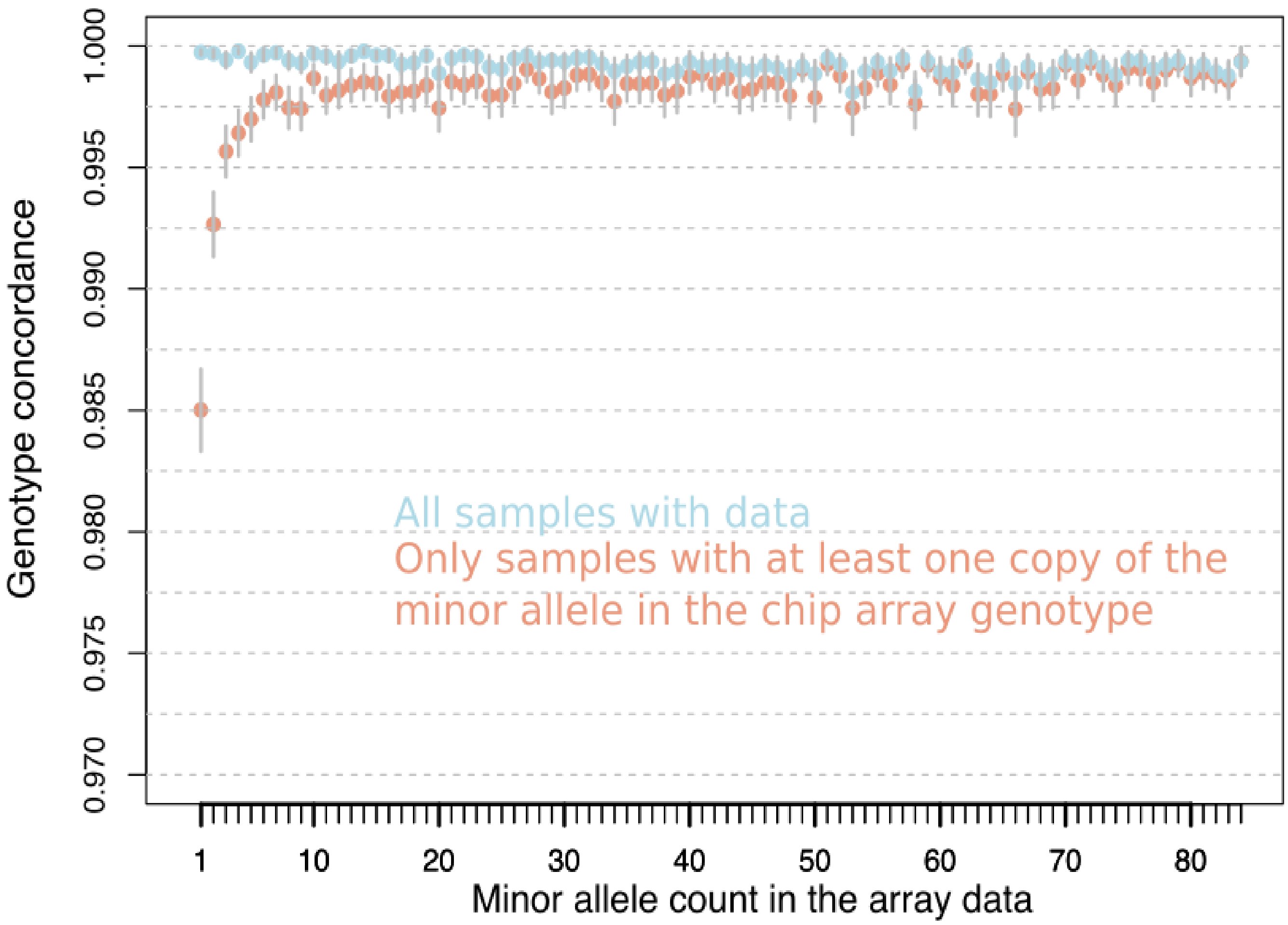
Many sites and remains dated to be older than 40 kya, suggesting a human settlement 47.5-55 kya

# One wave out of Africa *vs* Two waves out of Africa





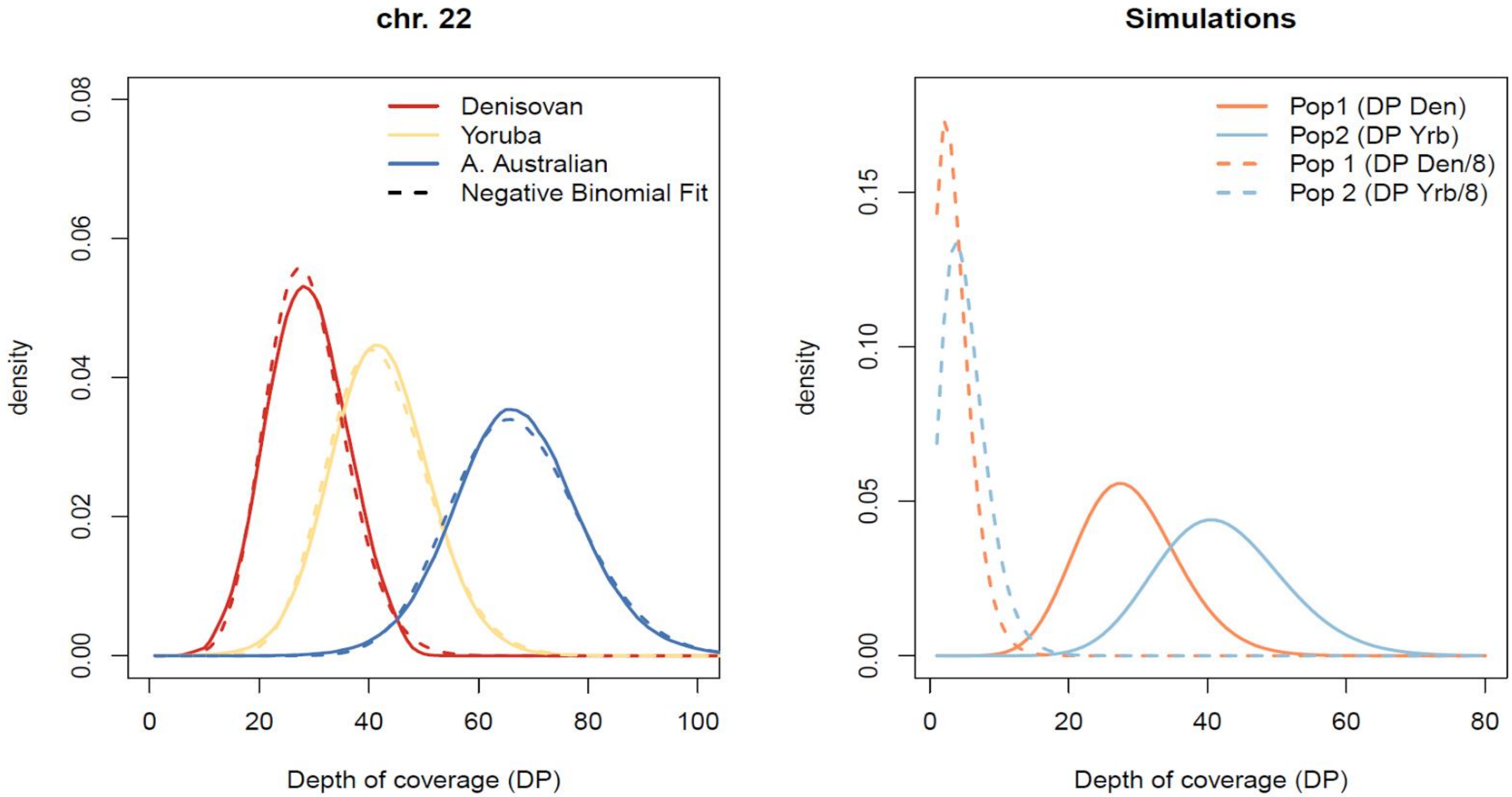
# 83 high-coverage Aboriginal Australians genomes



Average depth of coverage: 65x

Very good quality of genotype calls

# Effect of depth of coverage on SFS

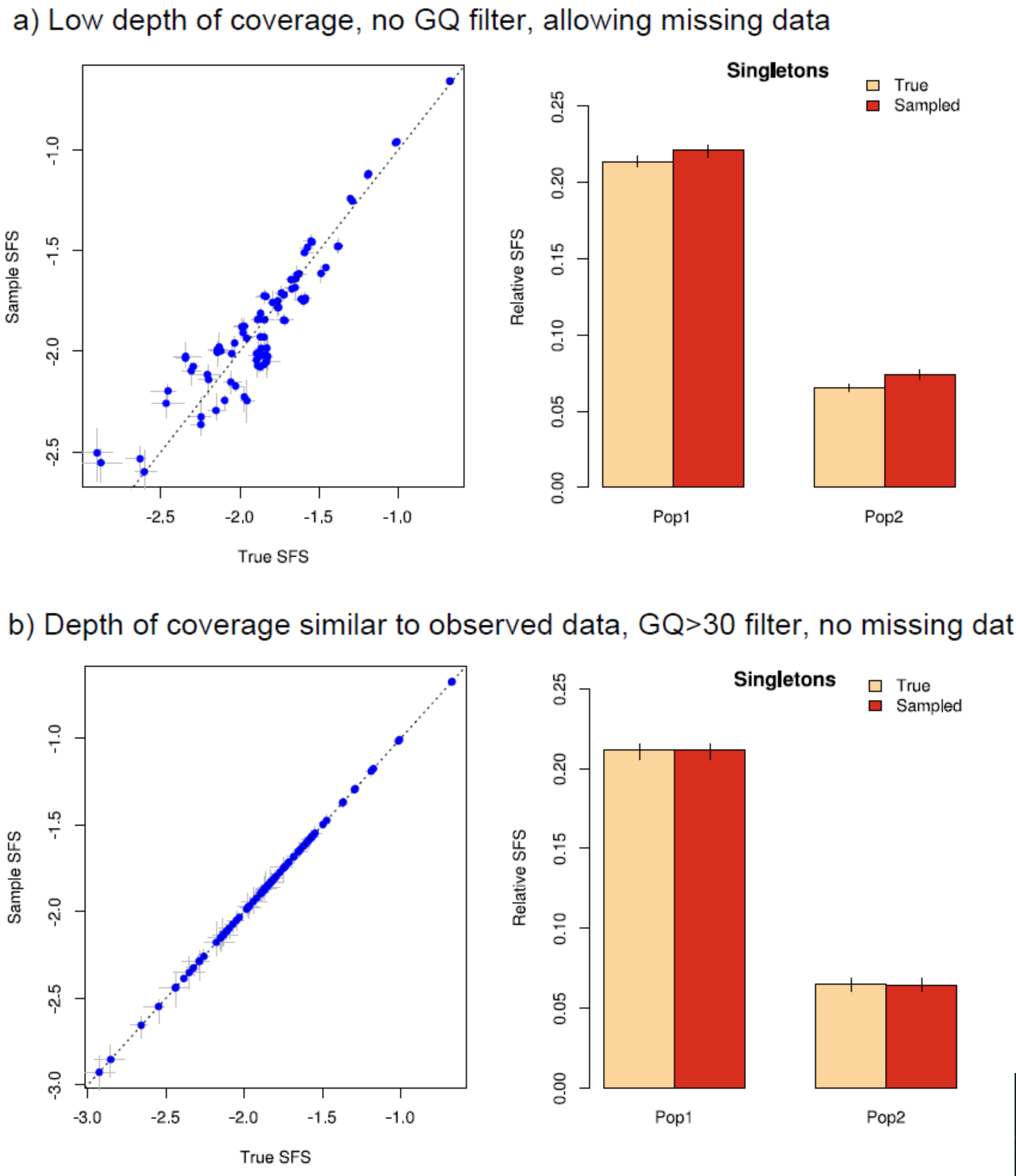


- Compared 2D SFS based on depth of coverage of observed data (mean larger than >20x), with a distribution 8 times smaller.

Malaspinas et al. (2016) Nature

# A note on recovering the SFS from genomic data

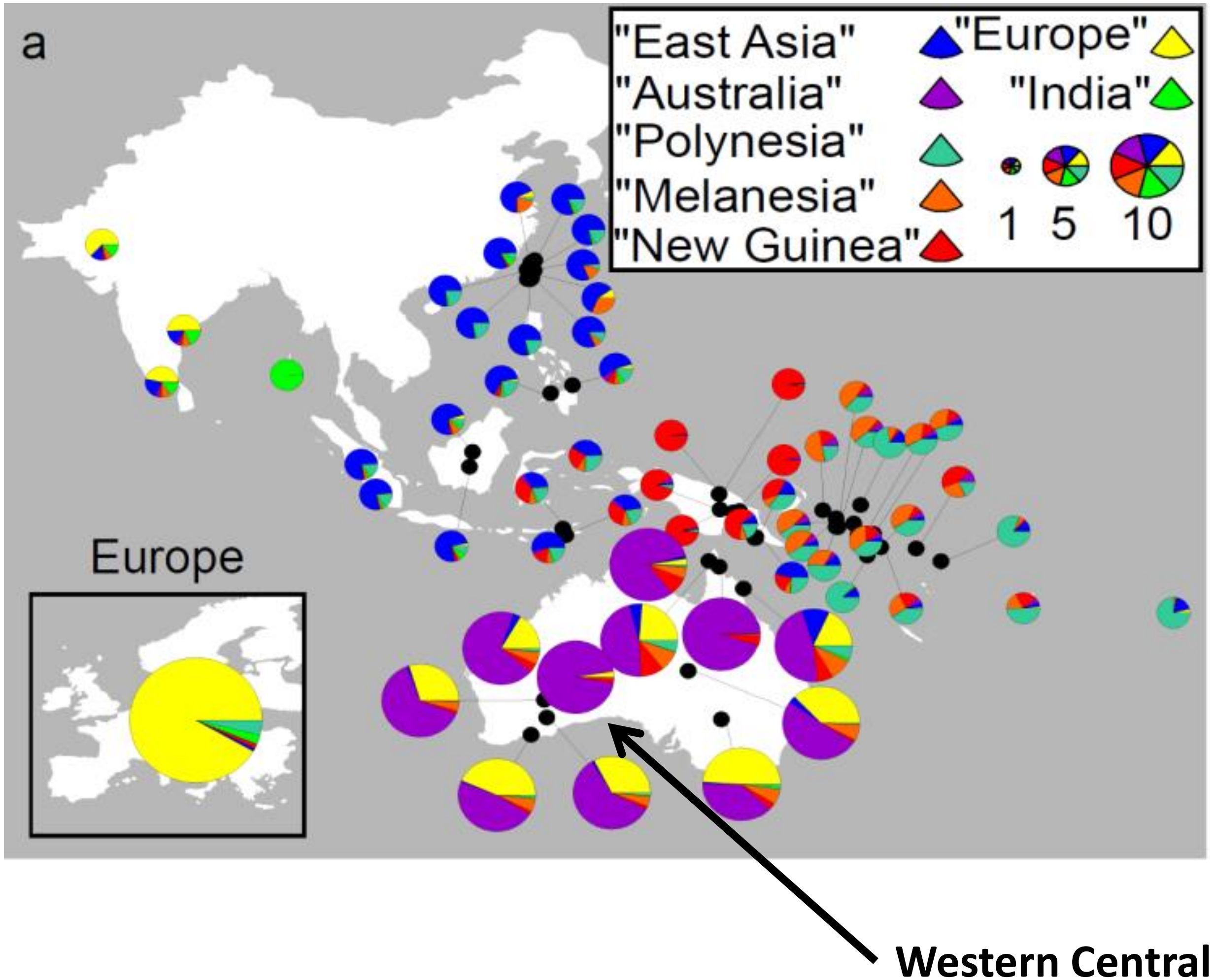
- Simulation study
- Low depth of coverage and missing data lead to biased SFS towards rare variants



a) Low depth of coverage, no GQ filter, allowing missing data

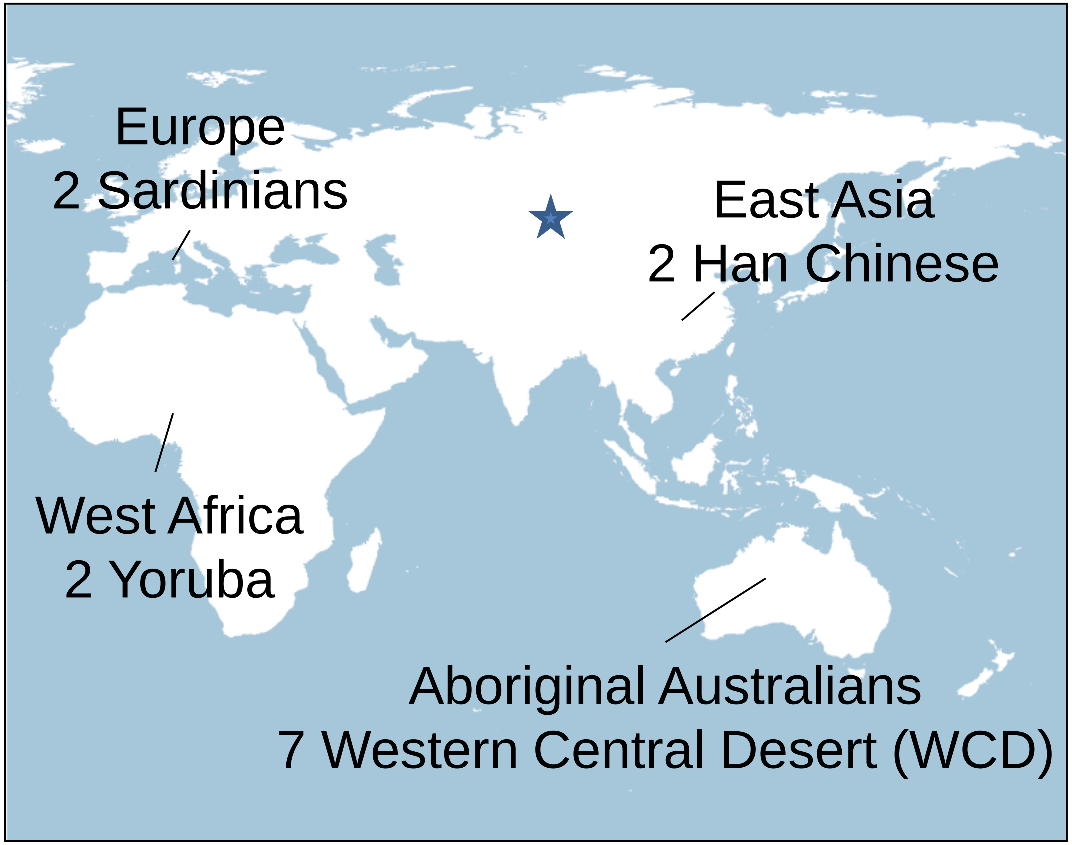
b) Depth of coverage similar to observed data, GQ>30 filter, no missing dat

# 83 high-coverage Aboriginal Australians genomes



**Western Central Desert (WCD)**

Average depth of coverage: 65x



★ Archaic human genomes:

- 1 Neanderthal (~66 kya)
- 1 Denisovan (~52 kya)

**Mutation rate assumed**

1.25 x $10^{-8}$ /site/gen

Scally and Durbin (2012) *Nat. Rev. Genet.*

**Generation time**

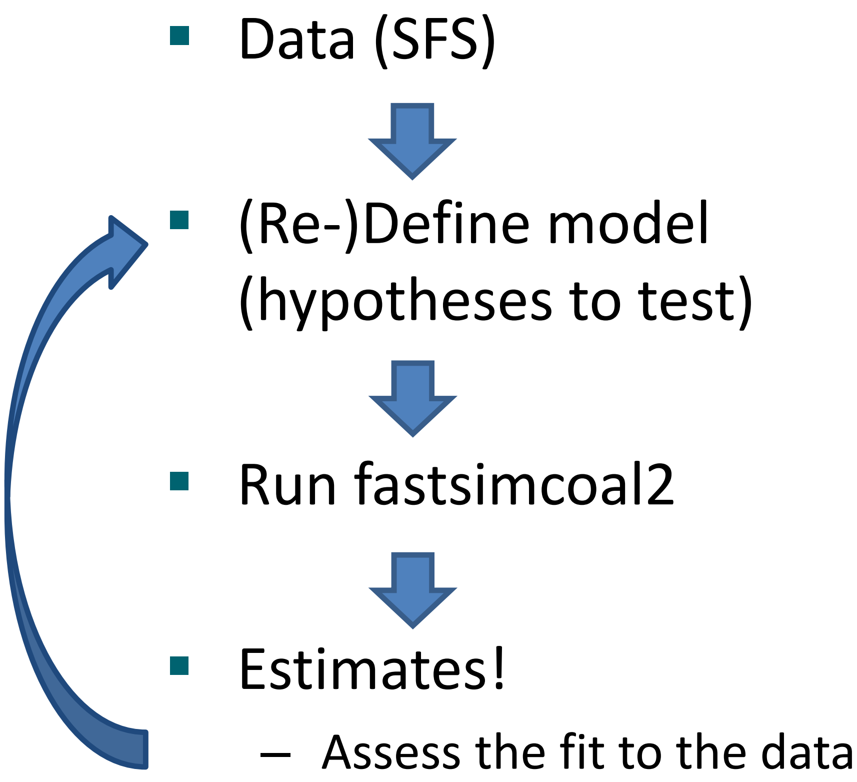
29 years/gen

Fenner (2005) *Am. J. Phys. Anthropol.*

Since we want to infer demography we tried to minimize the number of sites affected by selection:

- 985 1Mb blocks outside genic regions and CpG islands (~4.3 Million SNPs)
- 5 dimensional SFS (16,875 entries)
- Confidence intervals obtained using block-bootstrap

# Towards a model to test the hypotheses:
# One *vs* Two waves Out of Africa



- Data (SFS)
- (Re-)Define model (hypotheses to test)
- Run fastsimcoal2
- Estimates!
  - Assess the fit to the data

**Do you have an outgroup?**

- **Yes** – use the derived (unfolded) SFS
- **No** – use the minor allele frequency spectrum (folded)

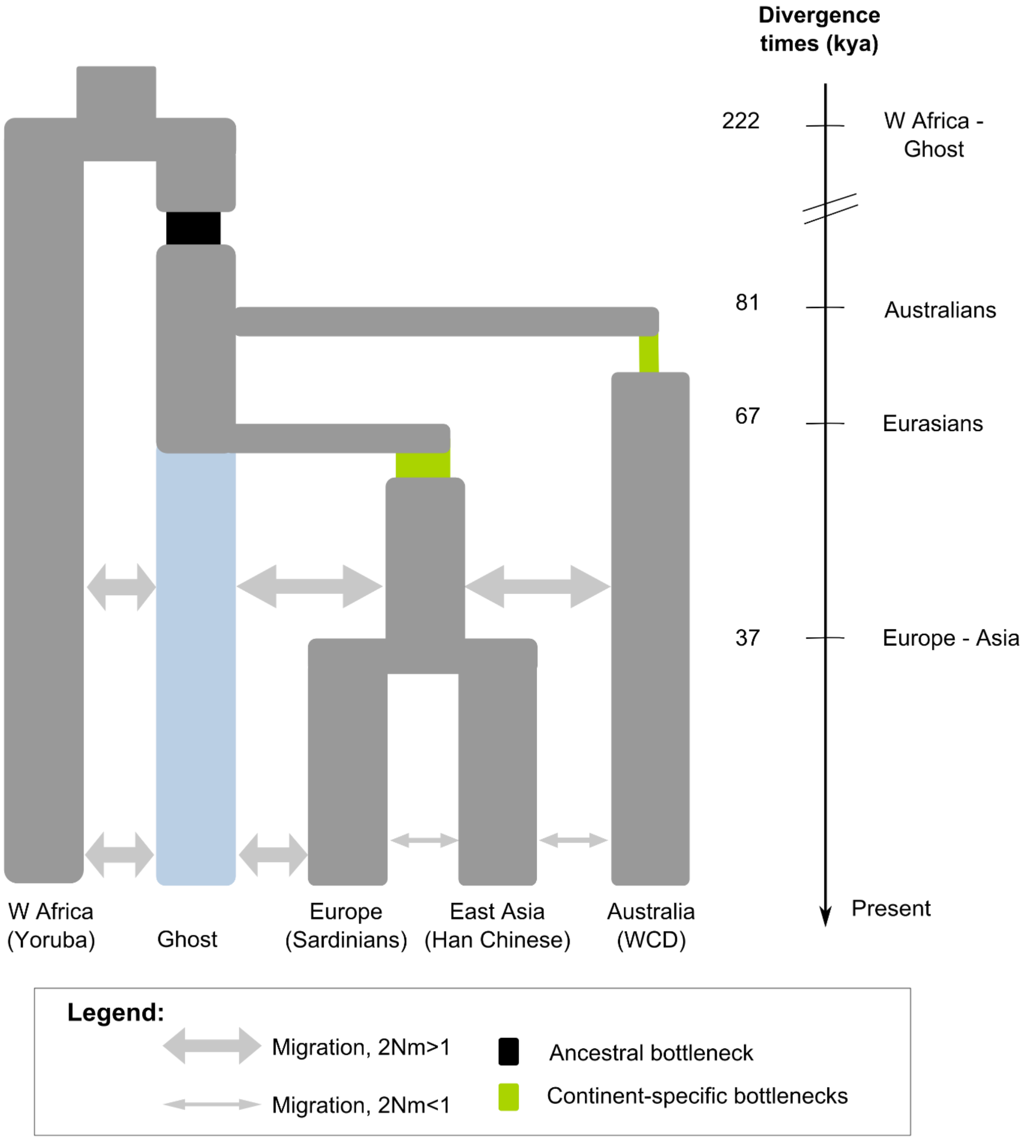
**Do you have monomorphic sites?**

- **Yes** - then, given a mutation rate you can infer the absolute times and effective sizes
- **No** – then all your estimates need to be relative to a fixed parameter (fixed Ne or fixed time)

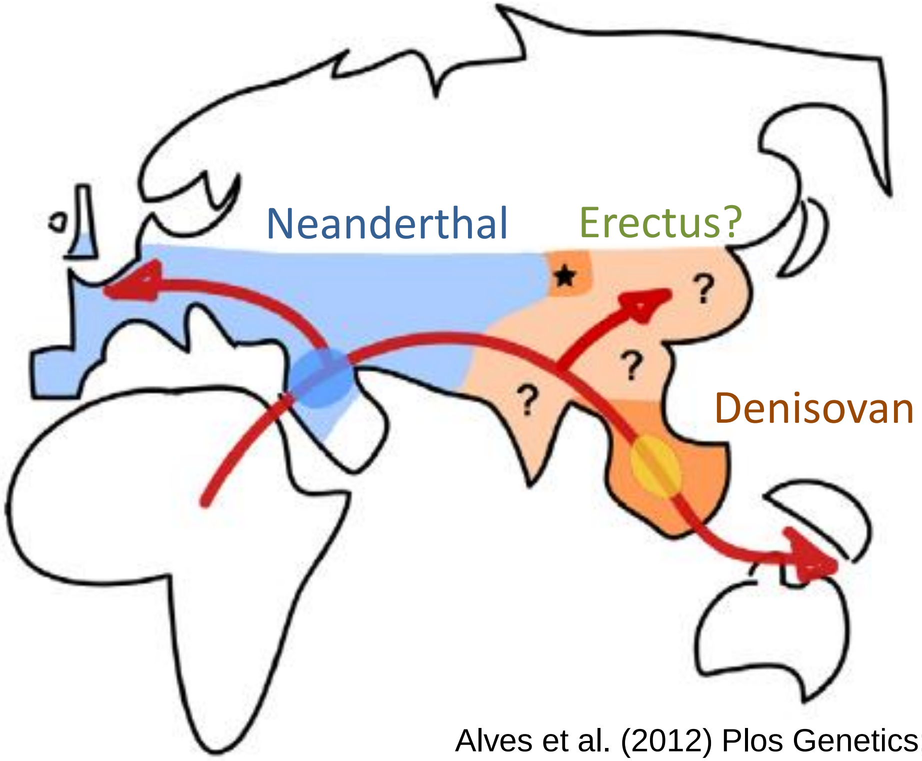
# We always get results…

## Evidence of two waves Out of Africa:

- Old split leading to colonization of Australia (81kya)
- More recent split leading to colonization of Eurasia (67 kya)



# Towards a model incorporating Neanderthal and Denisovan admixture

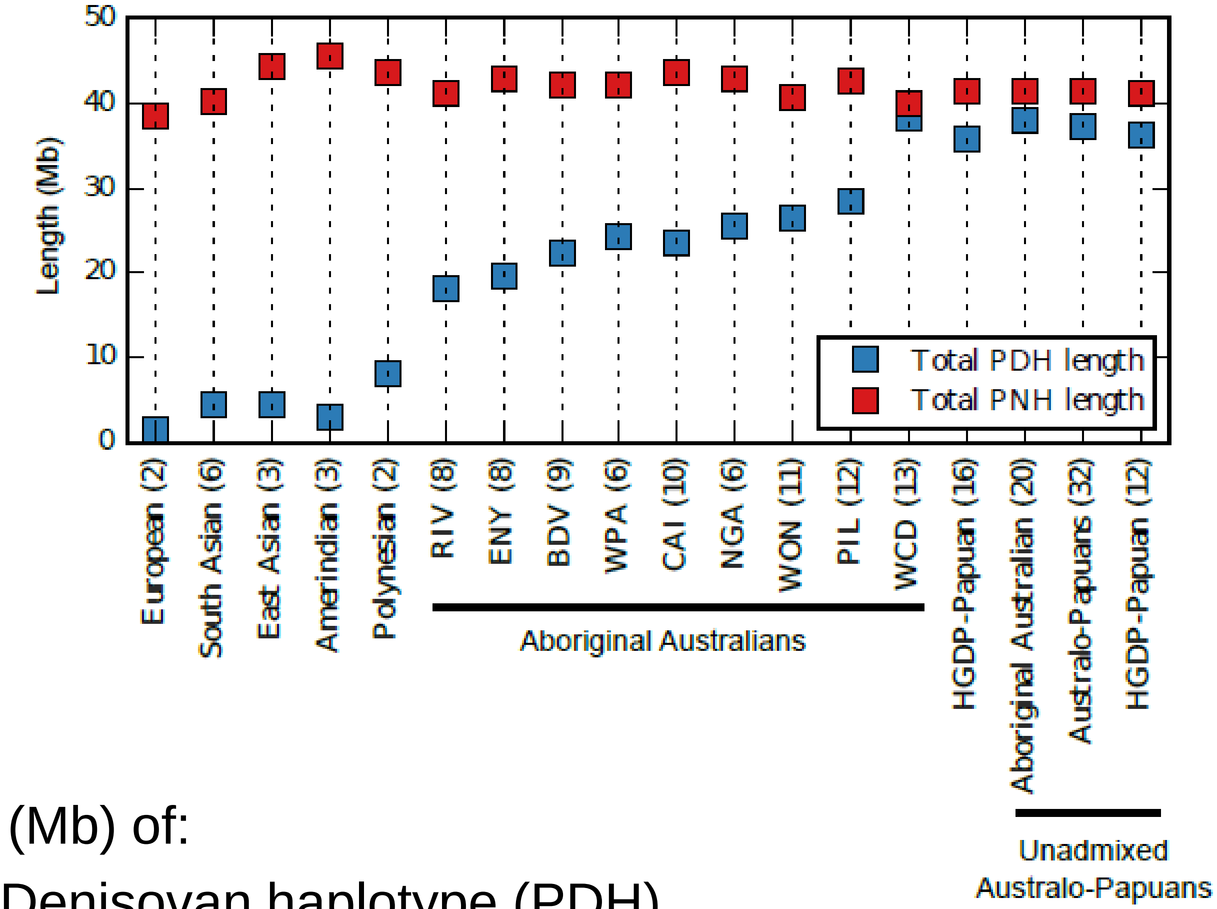


Alves et al. (2012) Plos Genetics;

- Non-African populations: 1-4% estimated Neanderthal admixture
- Aboriginal Australians and New Guineans: 3-6% estimated Denisovan admixture
- Archaic admixture can affect times of split estimates

Meyer et al. (2012) Nature; Prufer et al. (2014) Nature

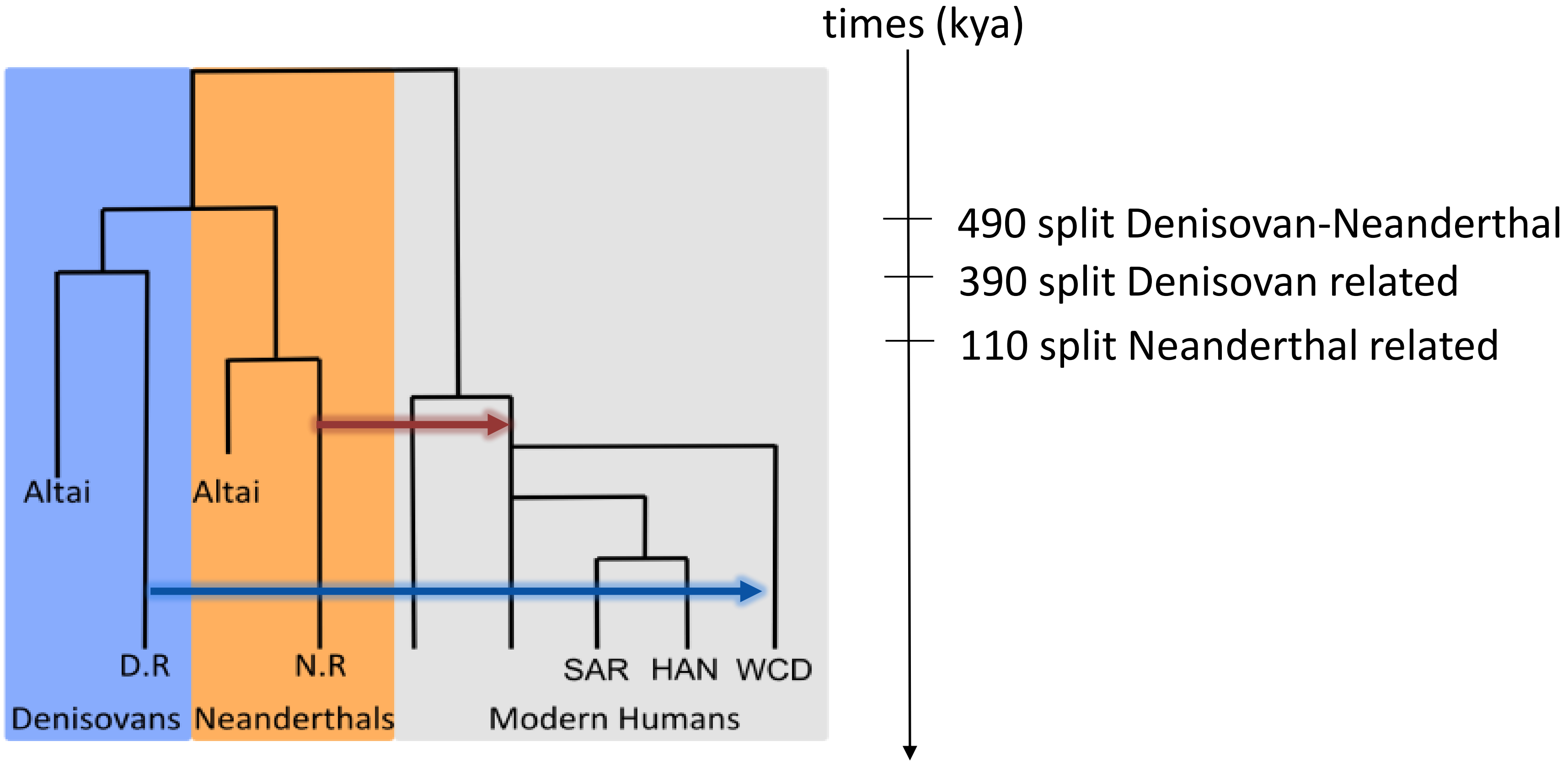
# Evidence of archaic introgression



Total length (Mb) of:

- Putative Denisovan haplotype (PDH)
- Putative Neanderthal haplotypes (PNH)

# Accounting for shared ancestry of Neanderthal and Denisovan



times (kya)

490 split Denisovan-Neanderthal

390 split Denisovan related

110 split Neanderthal related

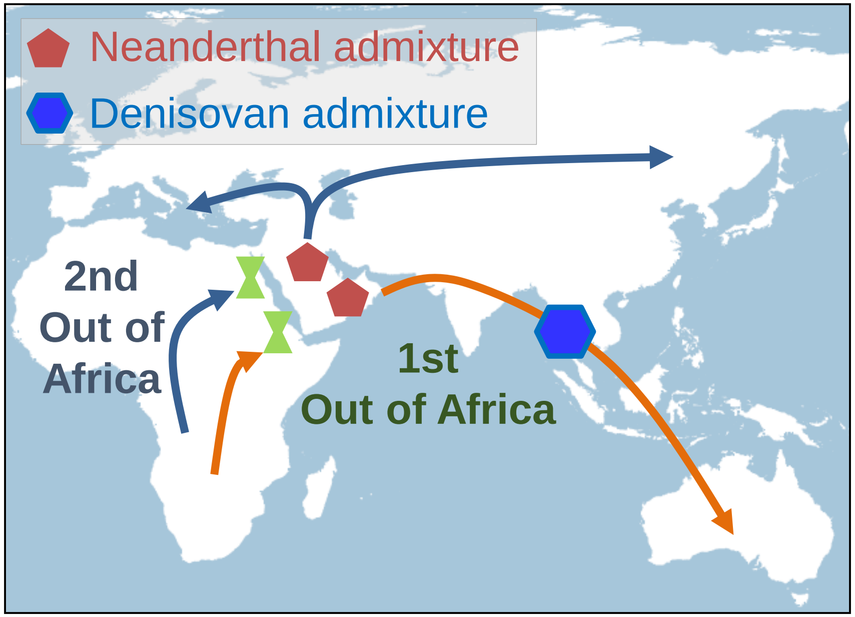
Admixture occurs between modern humans and:

– Denisovan-related (D.R.) population

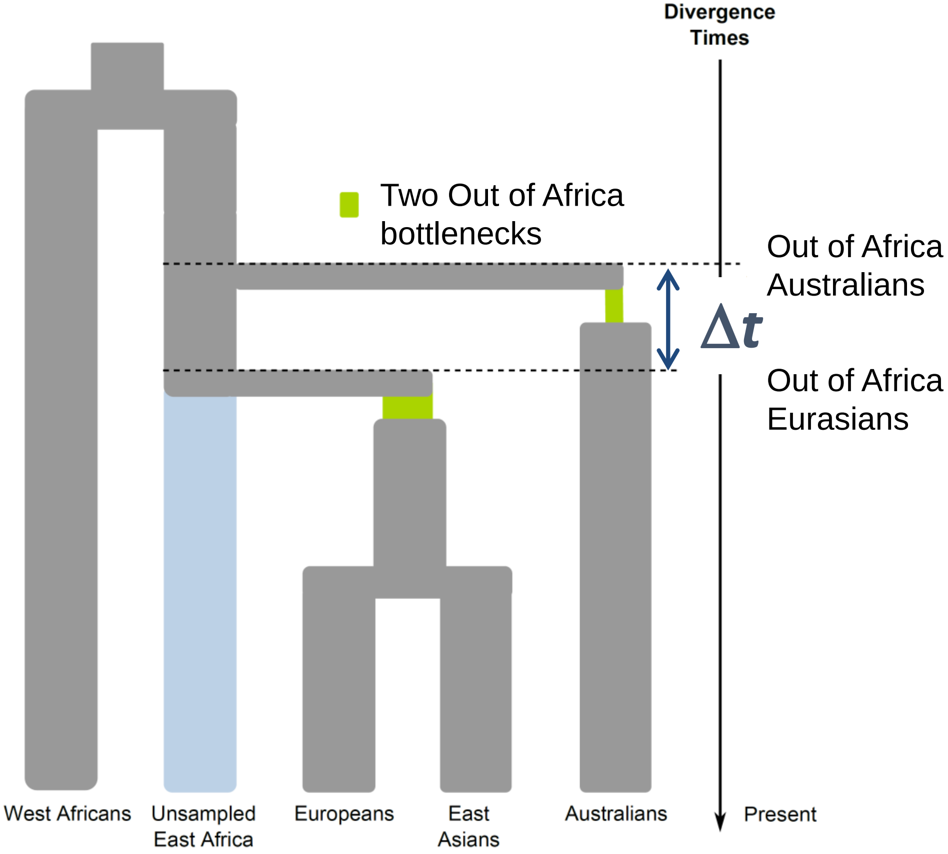
– Neanderthal-related (N.R.) population

Prüfer et al. (2014) Nature

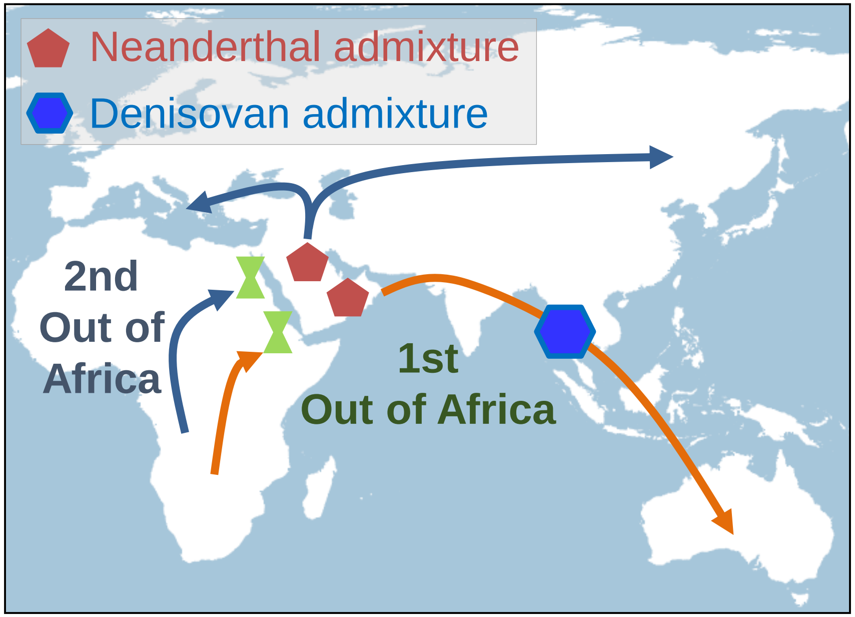
# Two-waves out of Africa



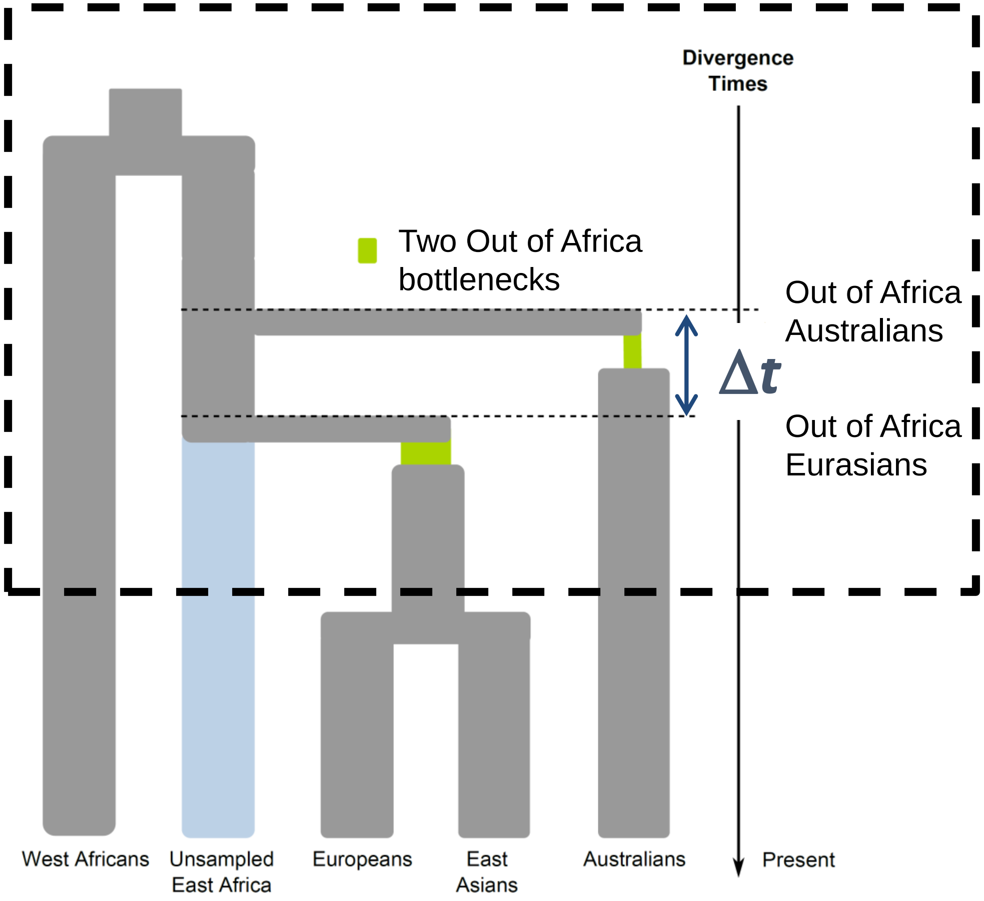
- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events



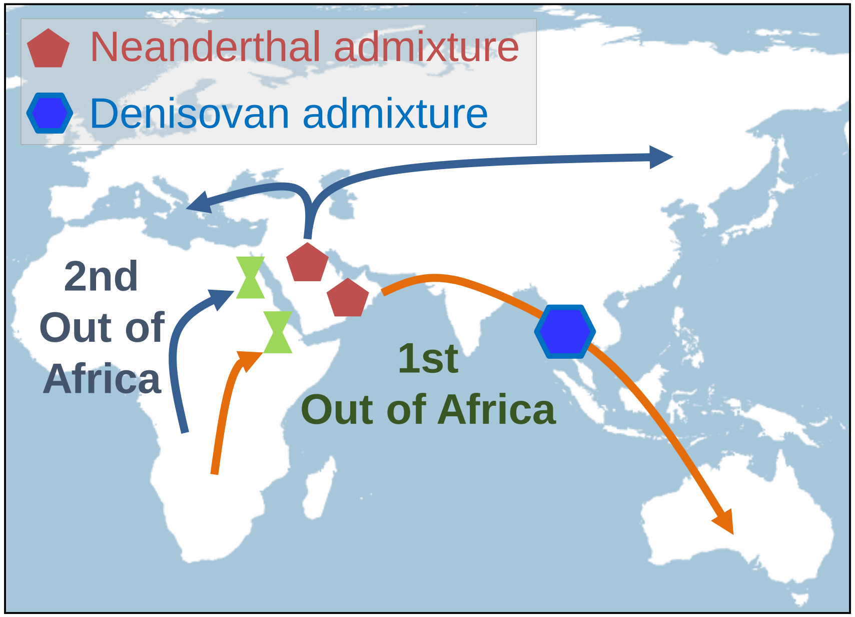
# Two-waves out of Africa

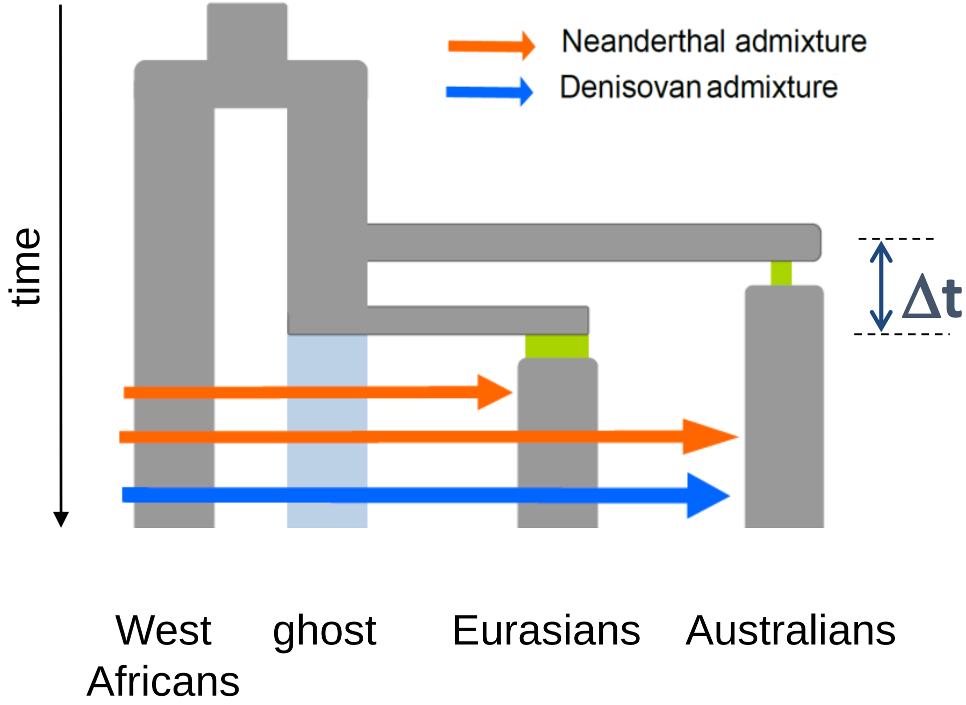


- Two different divergence times ($\Delta t >> 0$)
- Two independent bottlenecks associated with the two Out of Africa events



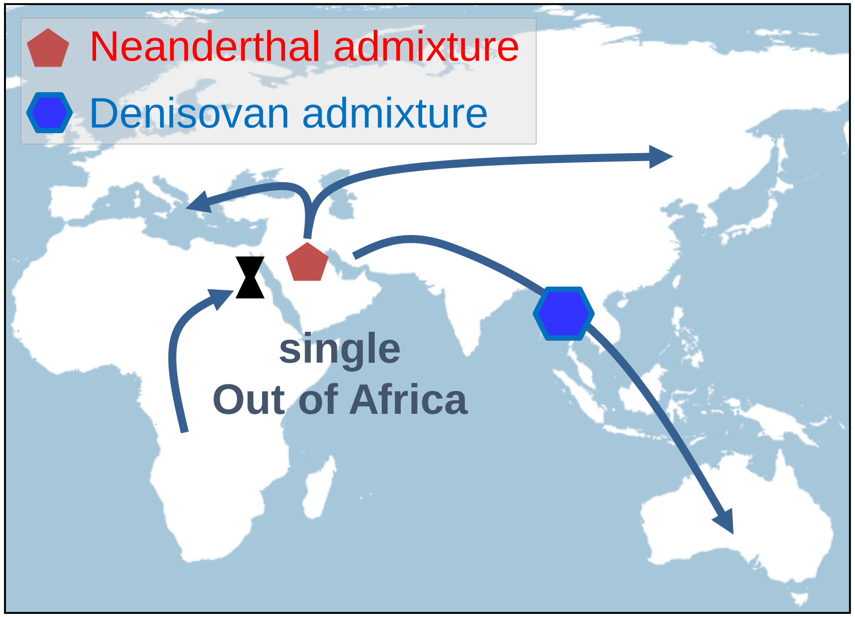
# Two-waves out of Africa

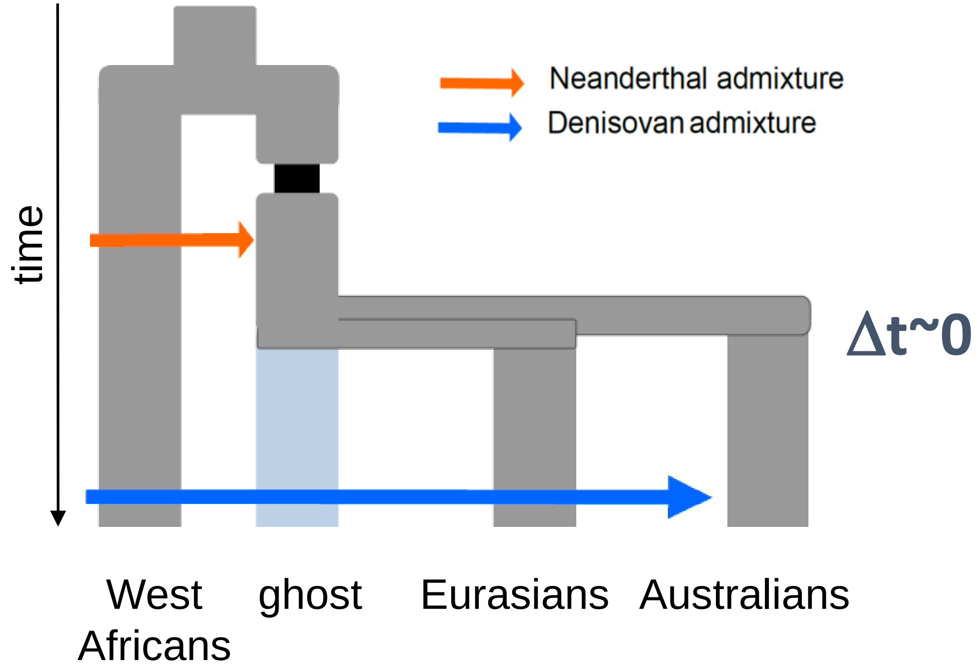




- Two different divergence times (Δt >> 0)
- Two independent bottlenecks associated with the two Out of Africa events

# One wave out of Africa

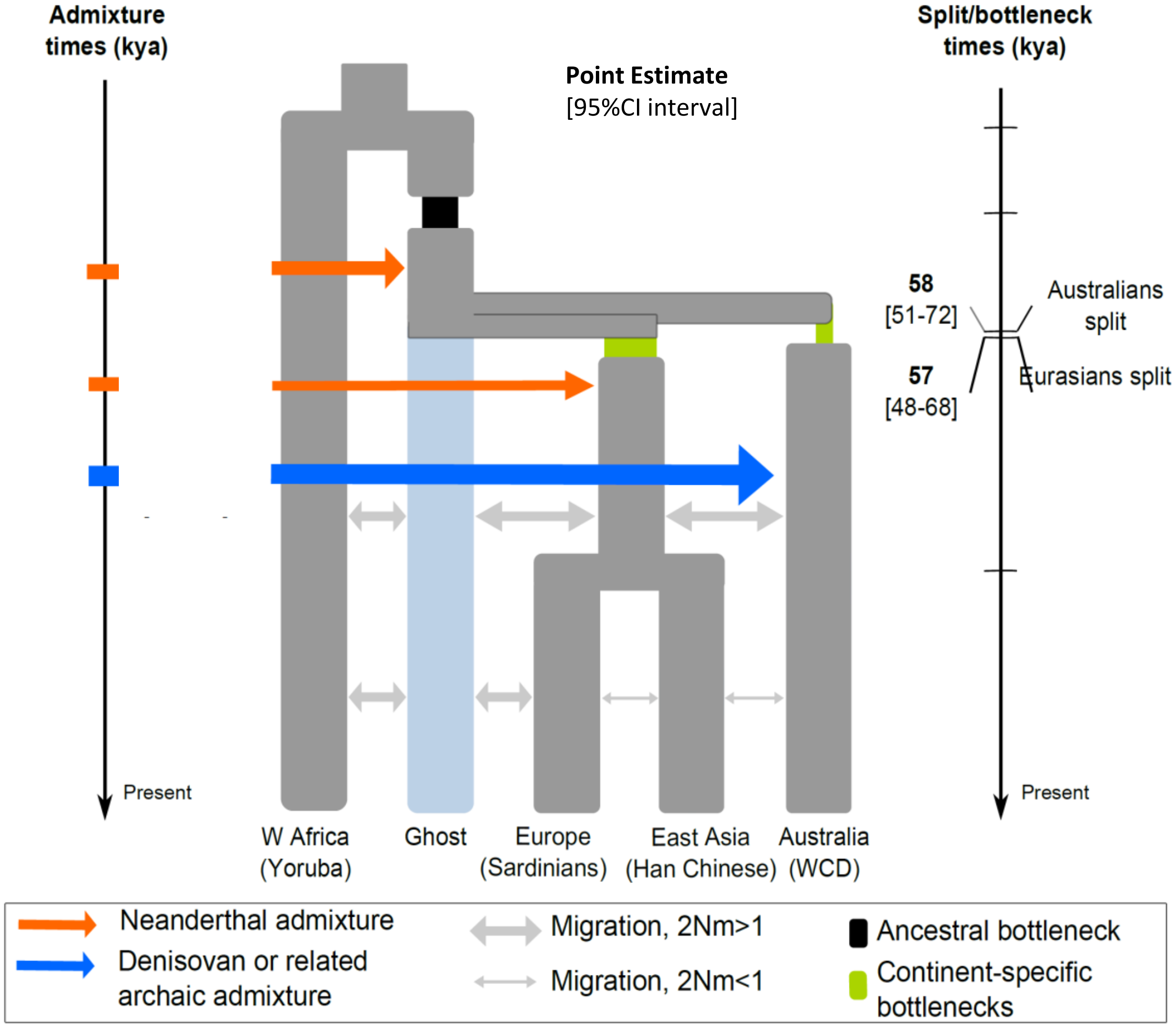




- Similar divergence times (Δt close to zero)
- One single bottlenecks associated with the Out of Africa events
- A major admixture pulse with Neanderthal

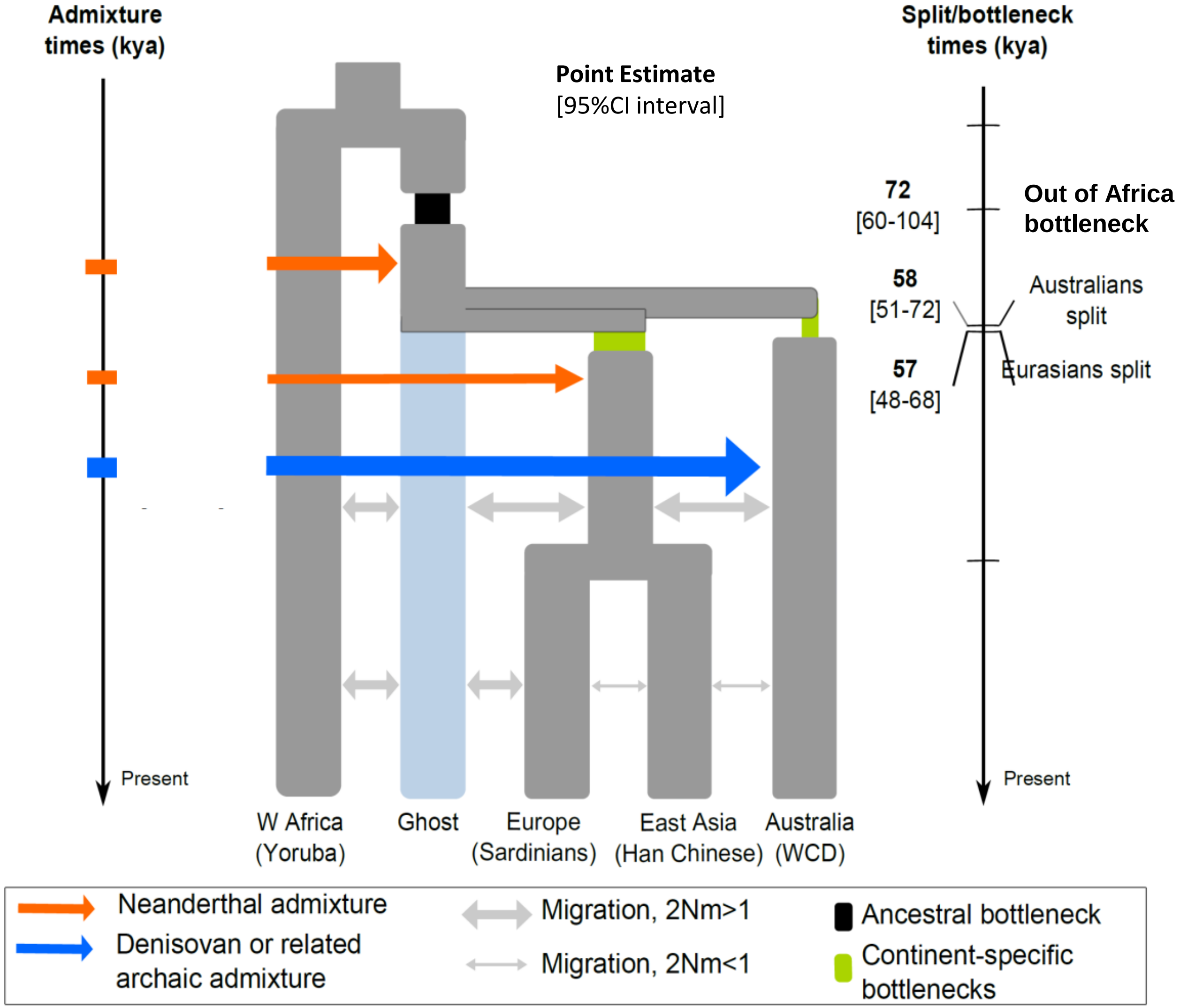
# A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)



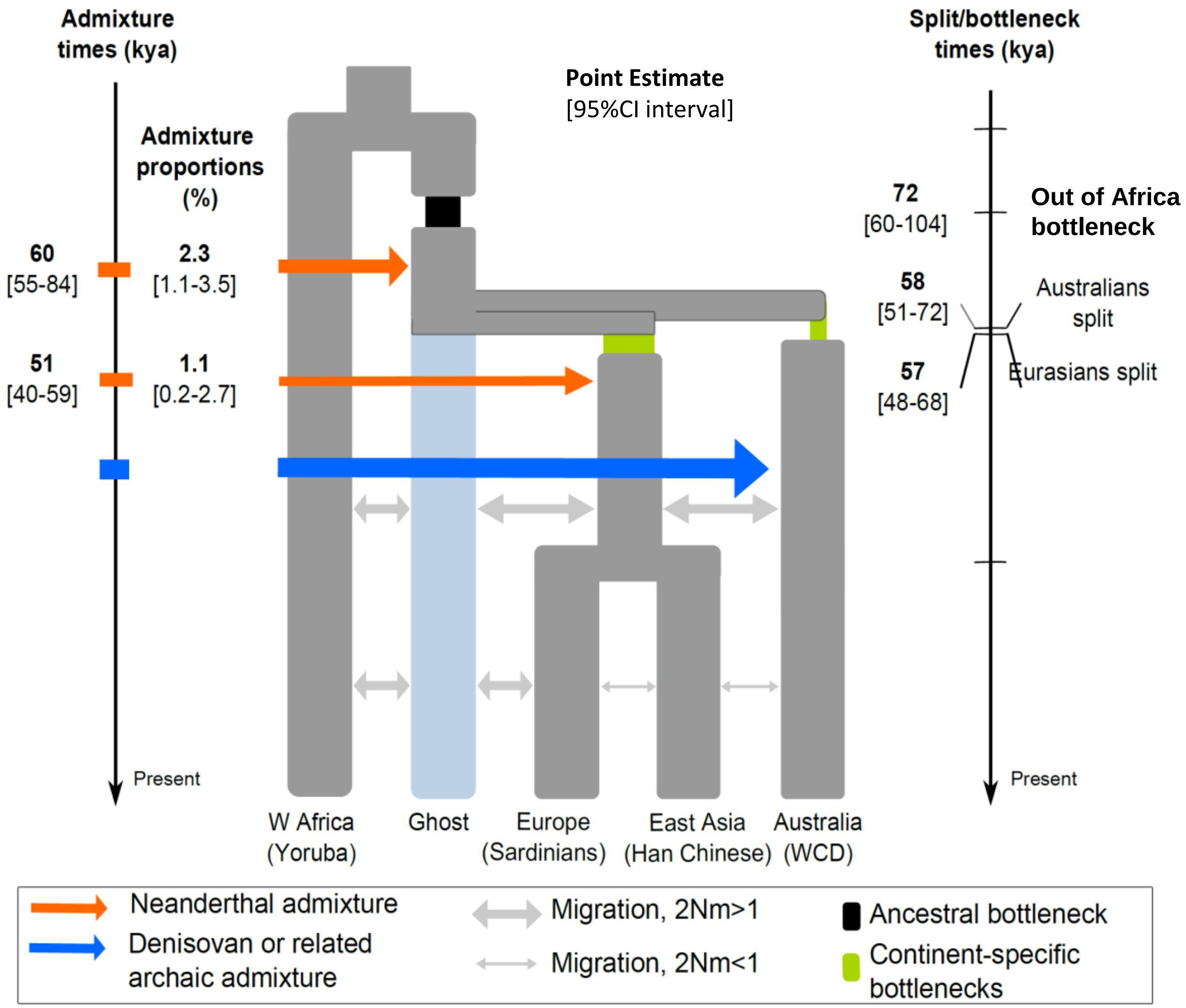
# A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event



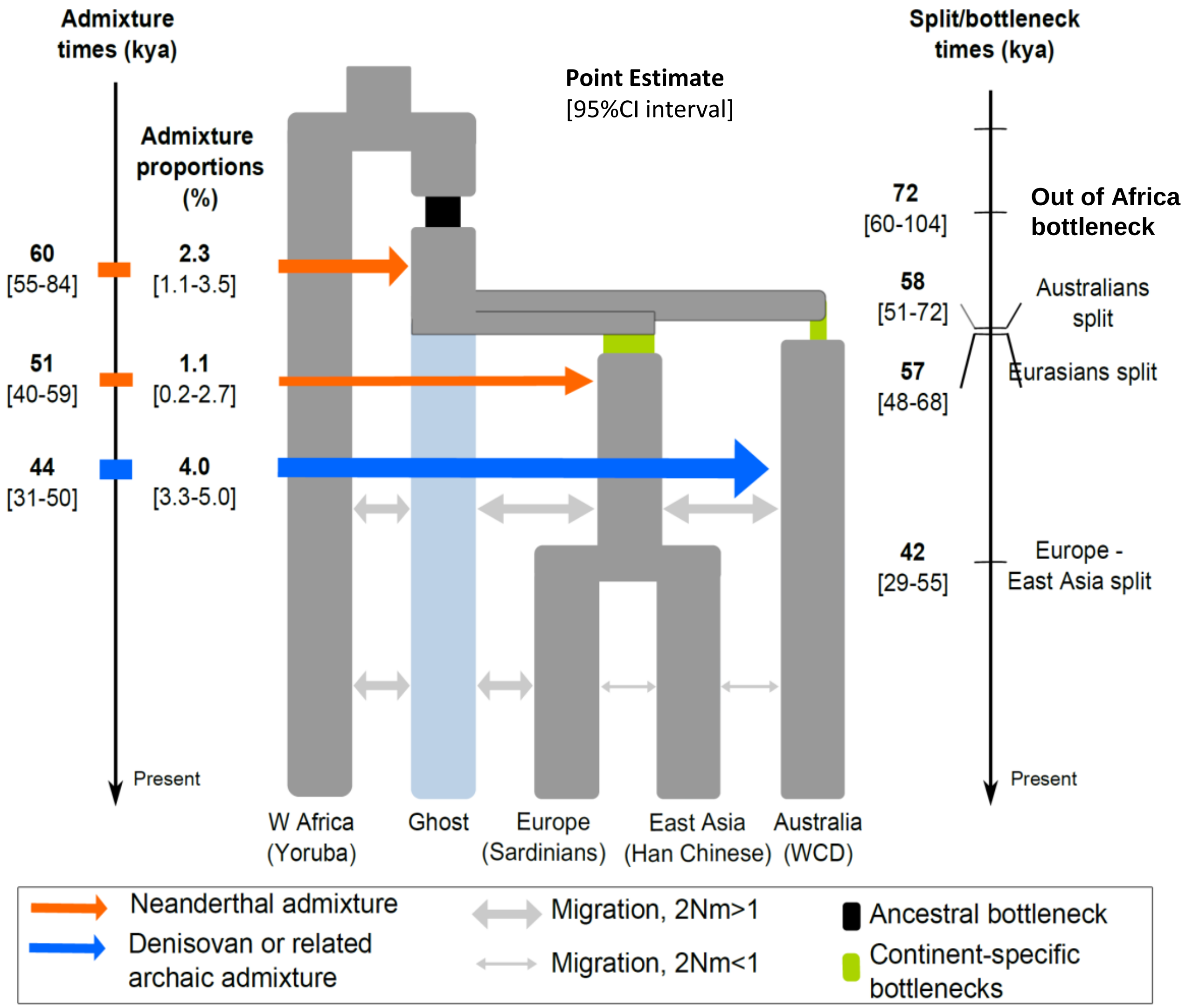
# A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans



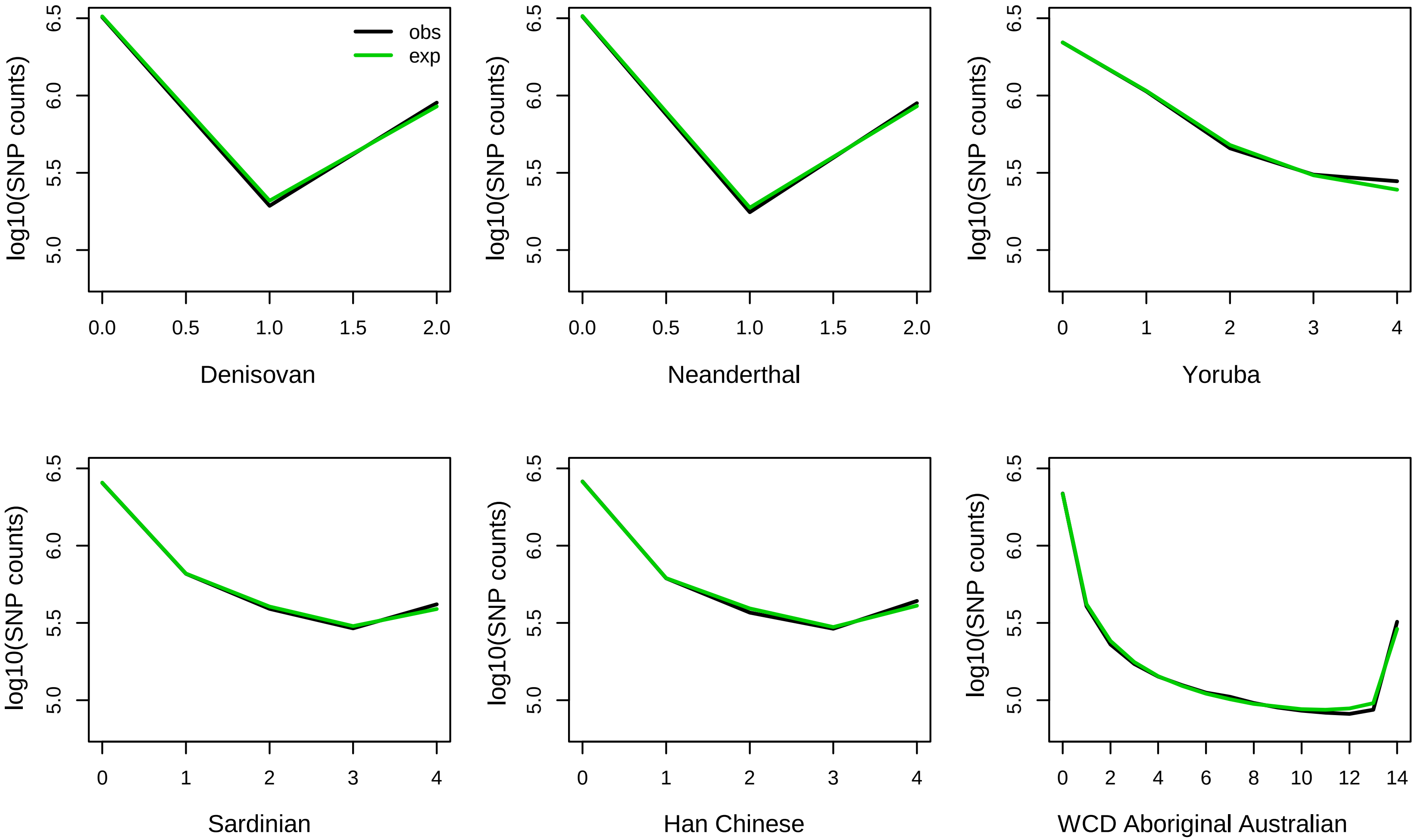
# A single wave Out of Africa is consistent with our estimates when accounting for archaic admixture

- Similar divergence time (Δt close to zero)
- Bottleneck associated with the Out of Africa event
- A major admixture pulse with Neanderthal in ancestors of all non-Africans

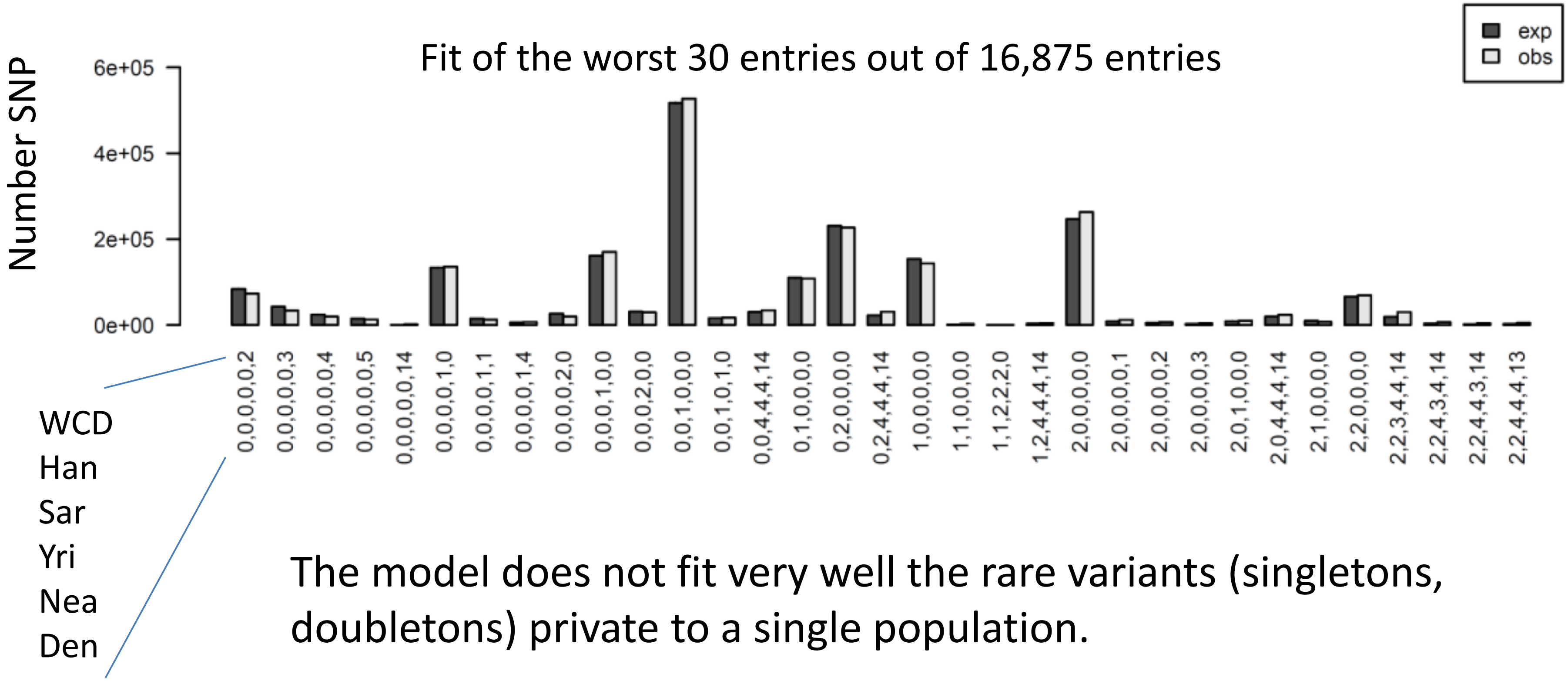


# Model captures aspects about the observed data

## Good fit to the marginal 1D site frequency spectrum



# What entries are not well fitted?



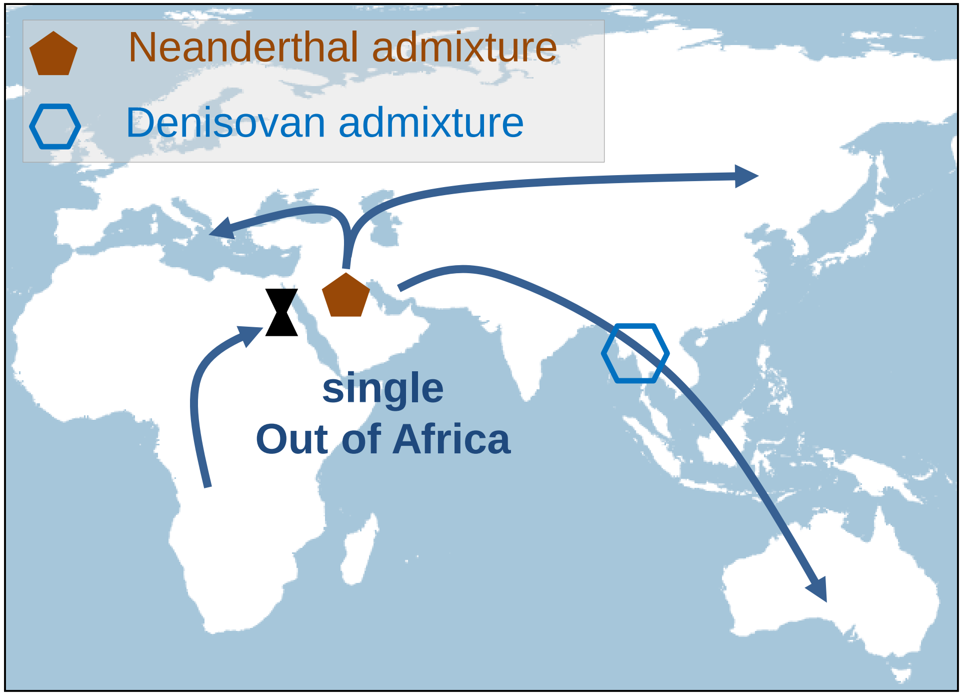
The model does not fit very well the rare variants (singletons, doubletons) private to a single population.

Pagani et al (2016) suggests two waves: Papuan genomes with signature of admixture with humans from first wave (at least 2% of their genome).

# Summary

## Aboriginal Australians genomes support a single major wave out of Africa

- Accounting for archaic admixture with Neanderthal and Denisovan was crucial to understand population divergence
- Genomic data consistent with a single major dispersal event out of Africa (60-104 kya)
- Two major dispersal waves into Asia: Aboriginal Australians diverged 51-72 kya from Eurasians

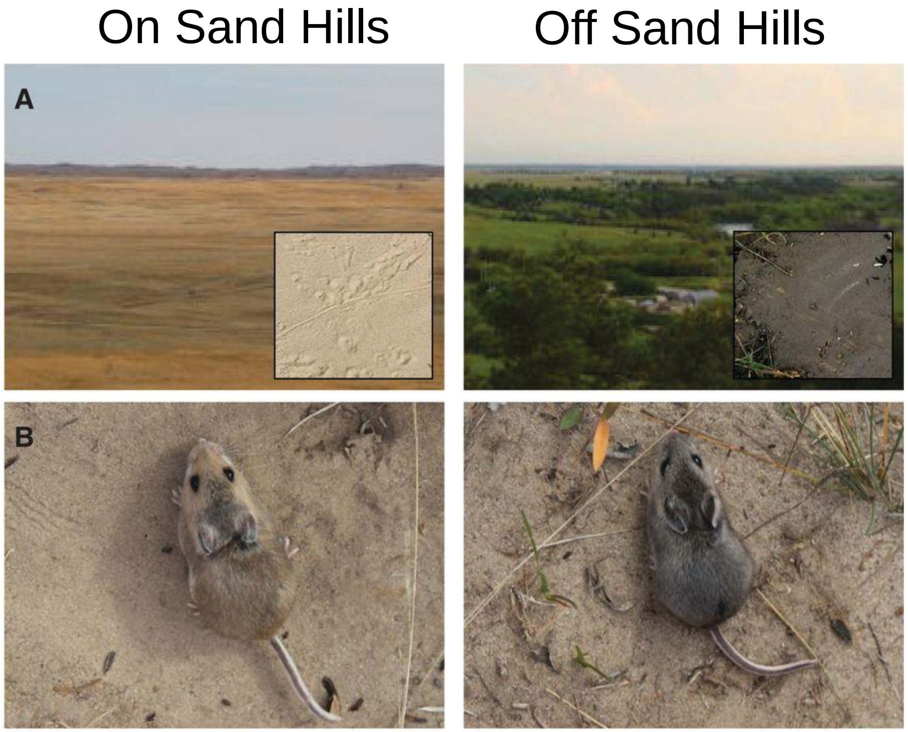


# Deer mice from Nebraska Sand Hills

S. Pfeifer, S. Laurent, V. Sousa, C. Linnen, H. Hoekstra, L. Excoffier, J. Jensen

# Coat color adaptation in deer mice *Peromyscus maniculatus*

- Habitat (soil color) correlated with coat phenotype
- Field experiments suggest that light color confers selective advantage against visually hunting predators
- Nebraska Sand Hills were formed 8000 to 15,000 years ago



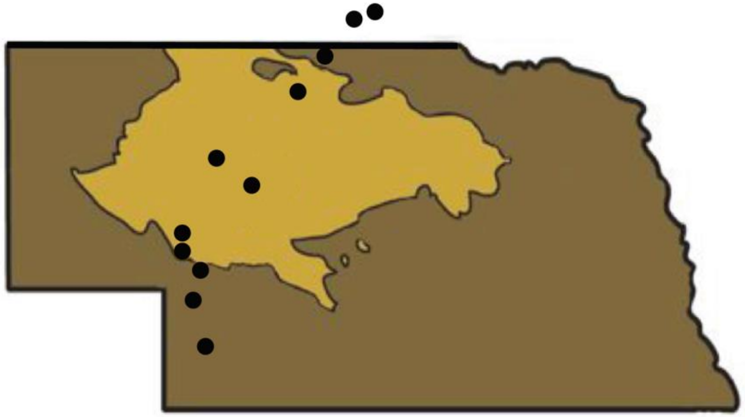
Linnen et al (2013) Science

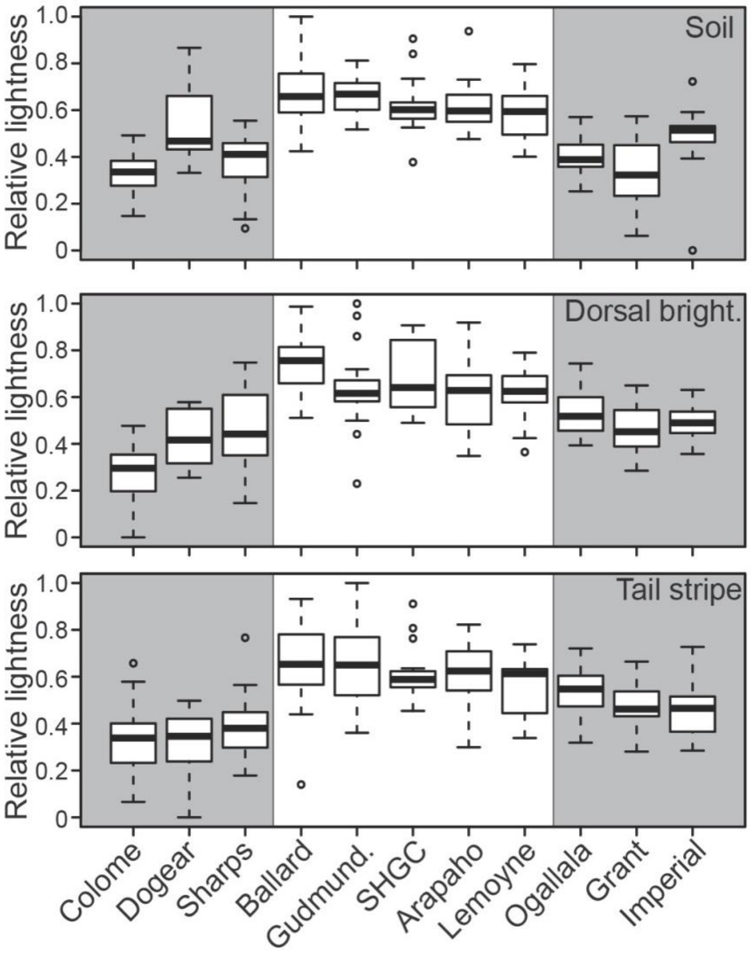
Pfeifer*, Laurent*, Sousa* et al (in press) MBE

# A transect across the Sand Hills (ON and OFF)

Sample locations “off” and “on” the Sand Hills

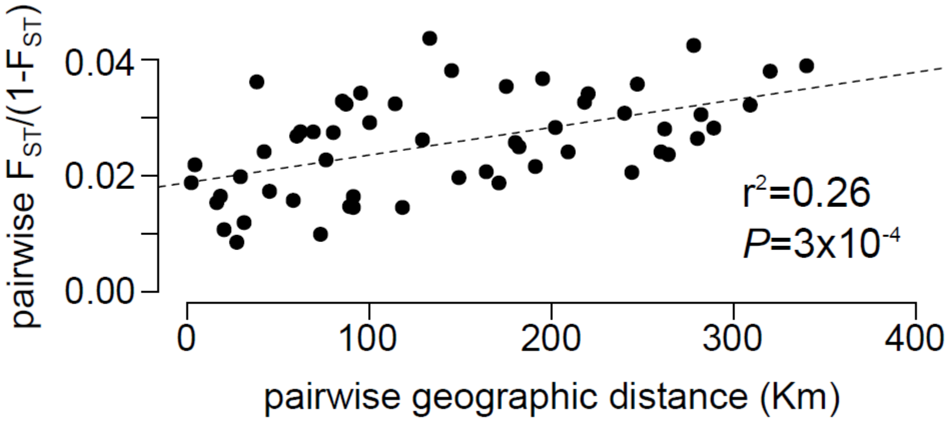
- 11 populations
- 330 individuals



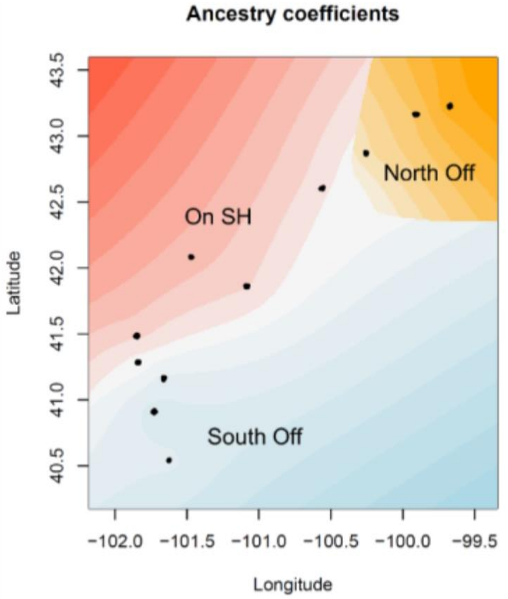


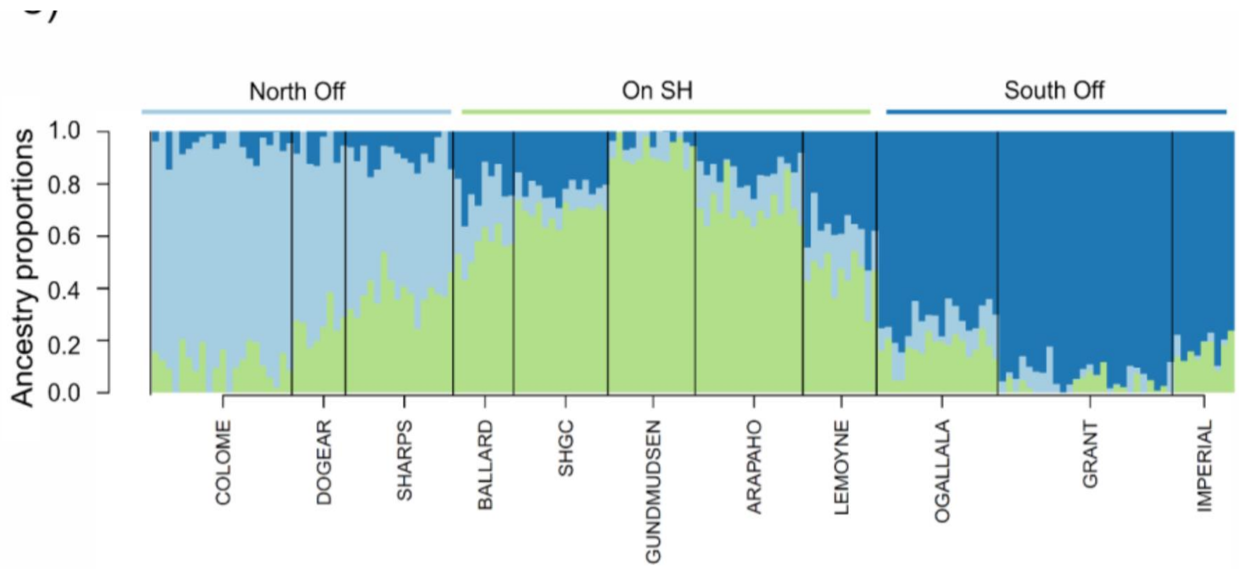
- Genomic data (NGS) data
  - Target 10,000 random 1.5kb regions
  - 185kbp region comprising the *Agouti* gene
- Phenotypic data for each individual

# Evidence for isolation by distance but three groups



Geographically closer samples are genetically more similar

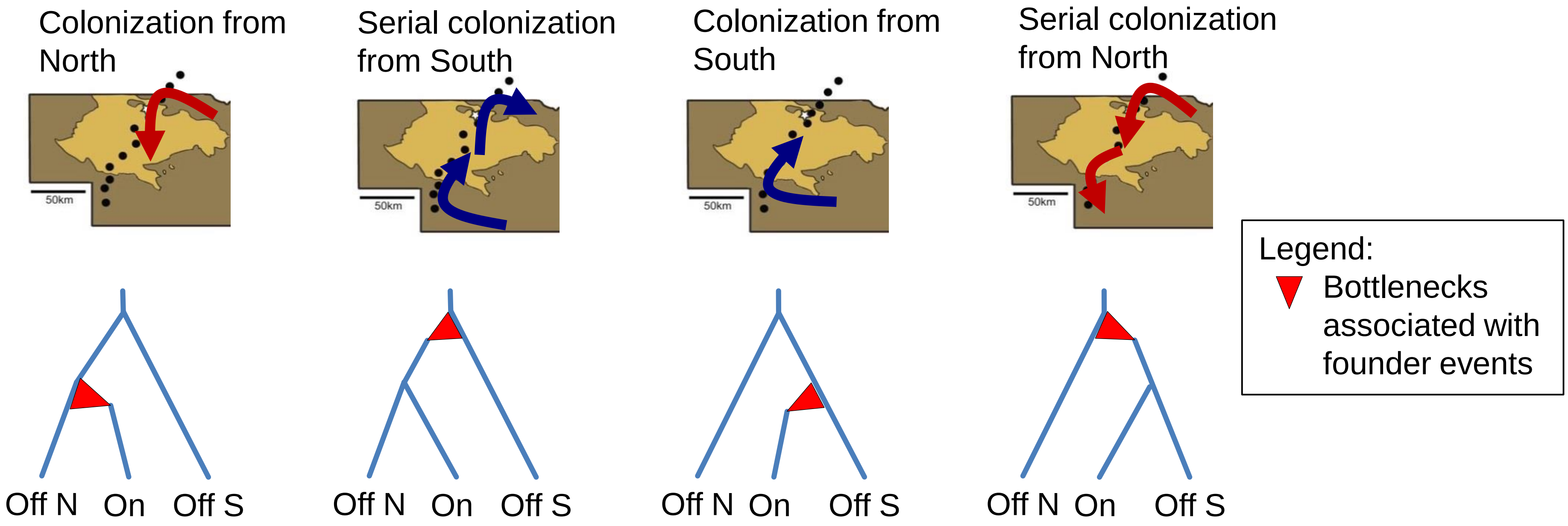




TESS3 analysis (ancestry estimation accounting for spatial information, Caye et al 2016)

# Model-based inference

Is there evidence of gene flow between Off and On the Sand Hills?



Colonization from North

Serial colonization from South

Colonization from South

Serial colonization from North

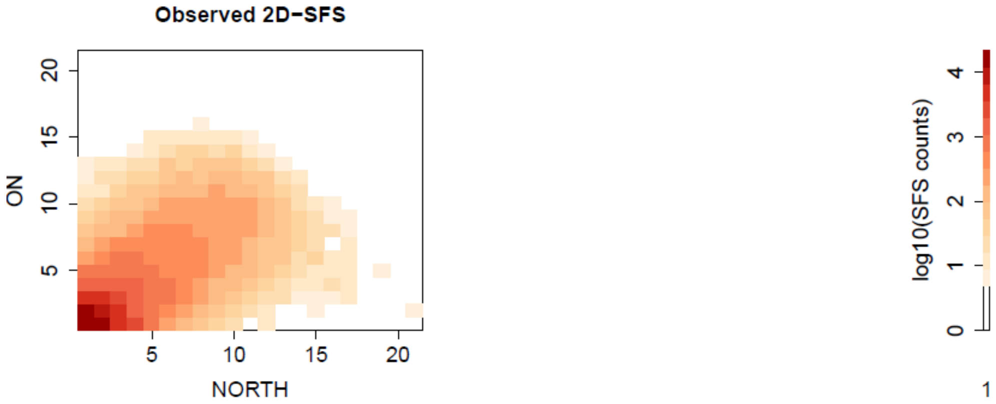
Legend:
▼ Bottlenecks associated with founder events

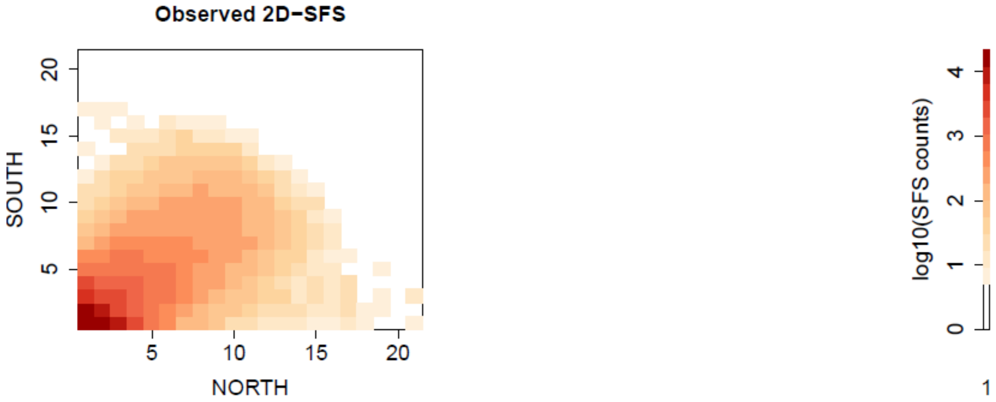
Estimates based on the joint **3D site frequency spectrum** (SFS):

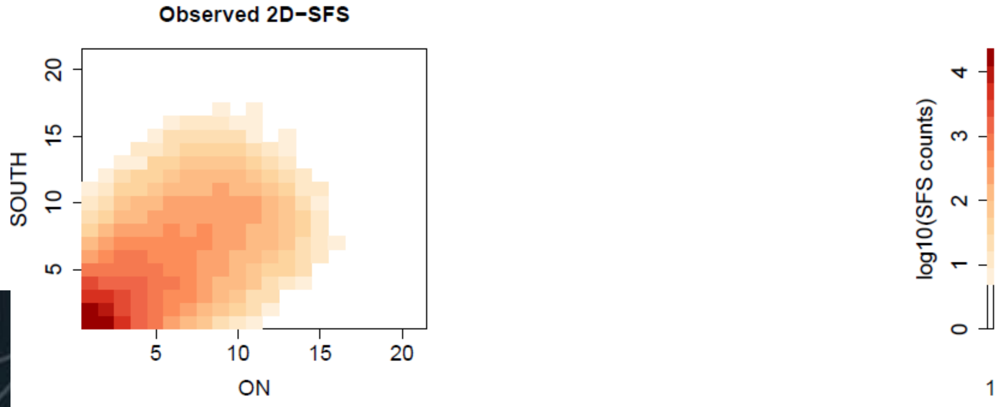
- folded SFS with 140,358 SNPs

Pooled individuals from three groups: north OFF, south OFF and ON the Sand Hills

# Deer mice: Pairwise marginal 2D SFS
Since we did not have an outgroup we used the folded SFS


Observed 2D-SFS

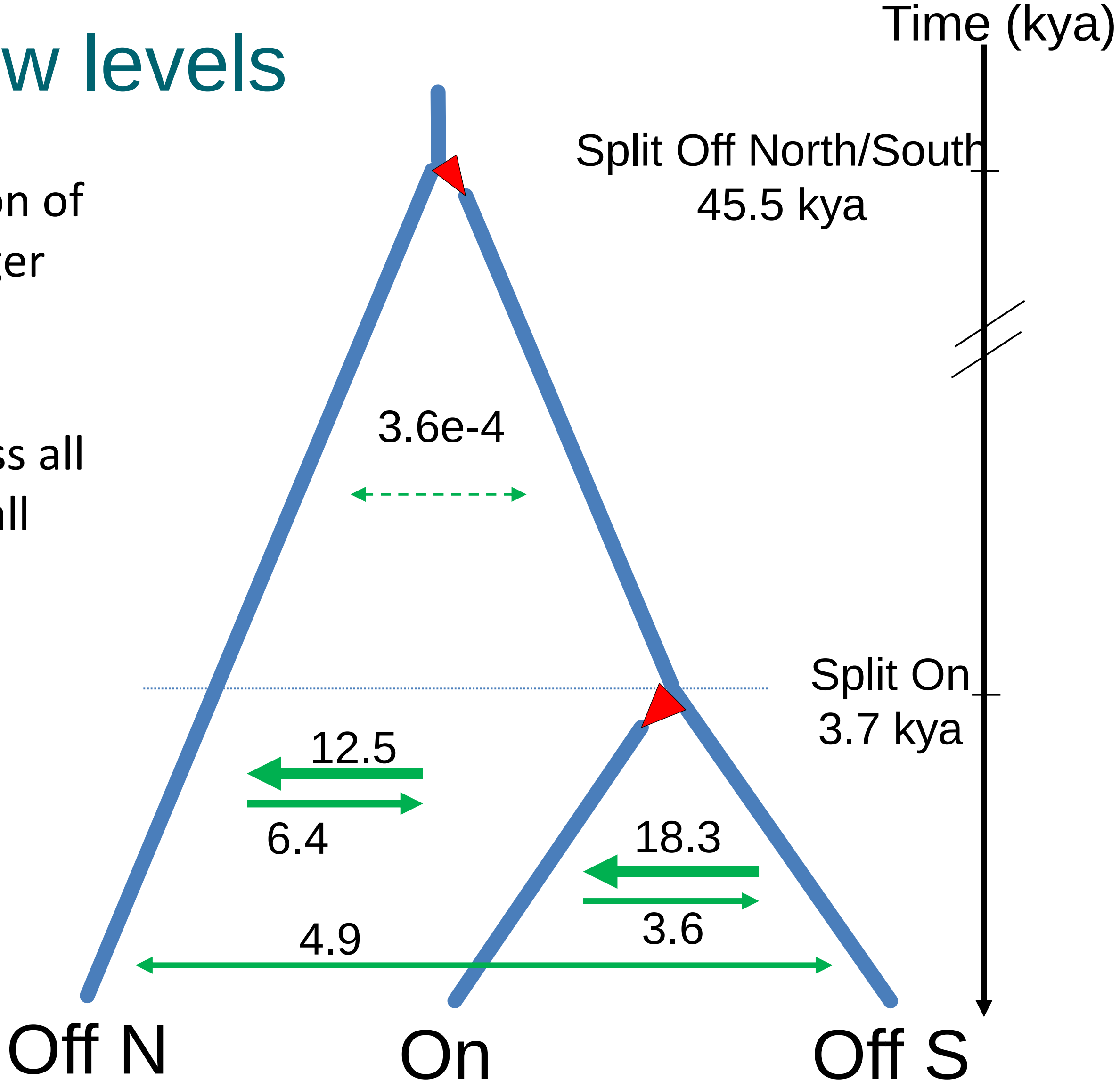

Observed 2D-SFS


Observed 2D-SFS

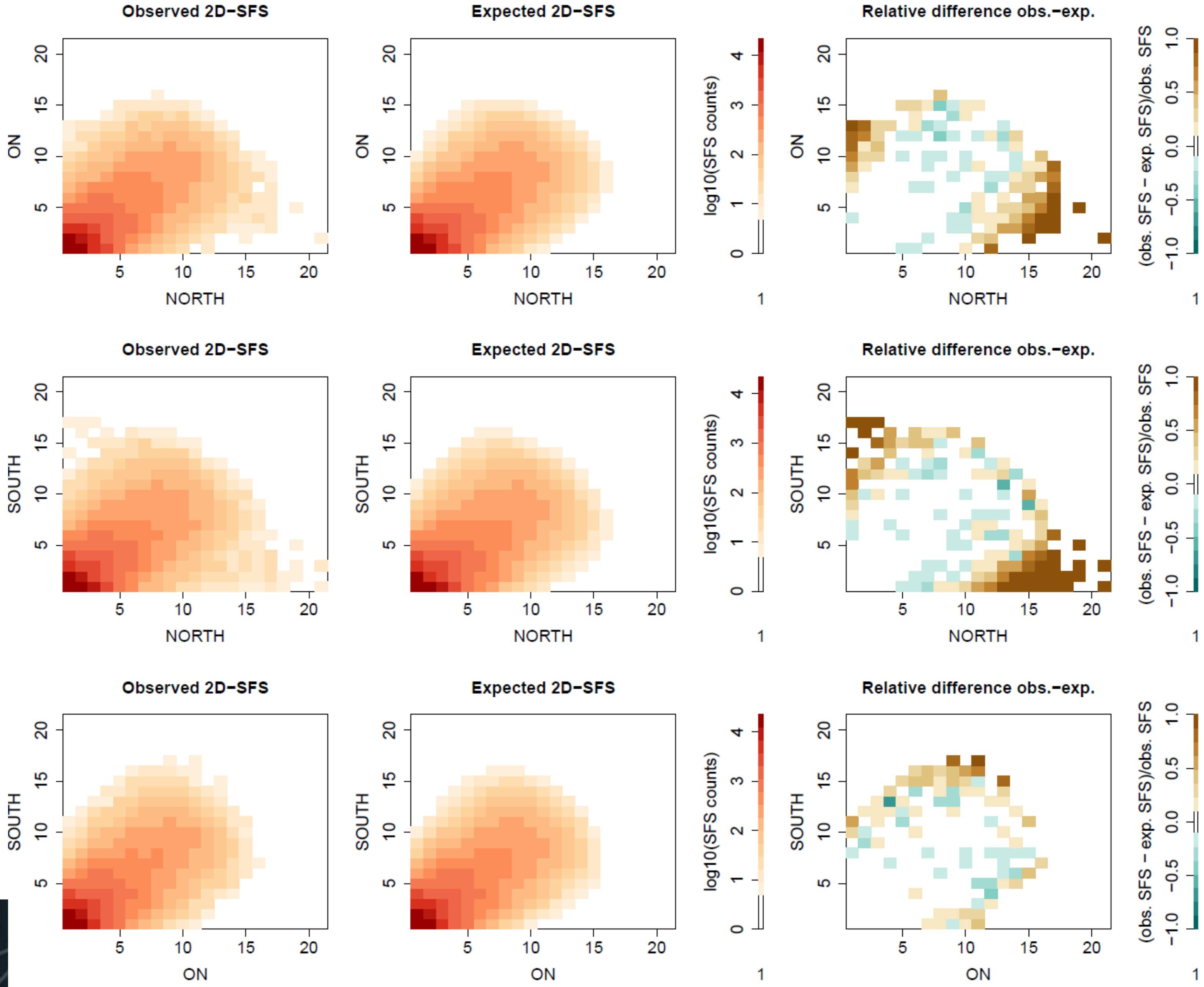
# Estimates support south colonization and high gene flow levels

- Recent time of colonization of Sand Hills ~3-5 kya, younger than formation of Sand Hills 8-15 kya
- High migration rates across all populations, inferred for all models

Migration rates above/below arrows in units of 2Nm, i.e. average number of immigrants per generation.



# Deer mice: Model fit to marginal SFS

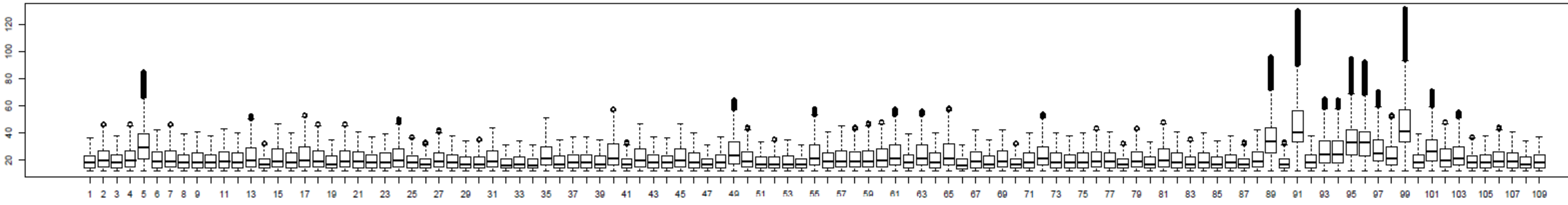


# Some lessons I learned working with the deer mice data

- Be carefull when applying Hardy-Weinberg filters to your data
- Be carefull when filtering on depth of coverage applying the same thresholds for all individuals

# The depth of coverage varied considerably across individuals

Example of the DP distribution for each individuals, for individuals with mean DP>12



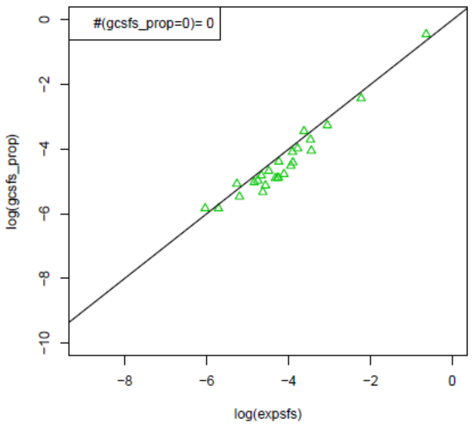
- Applying the same threshold for all individuals can lead to biases
- Apply a filter on DP for each individual

# Effect of DP filters on the SFS

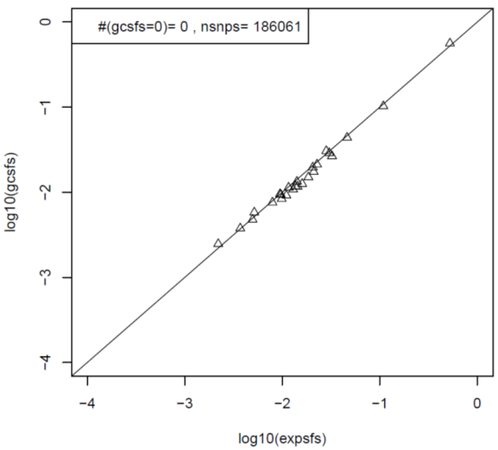
## Simulation study

SFS based on called genotypes

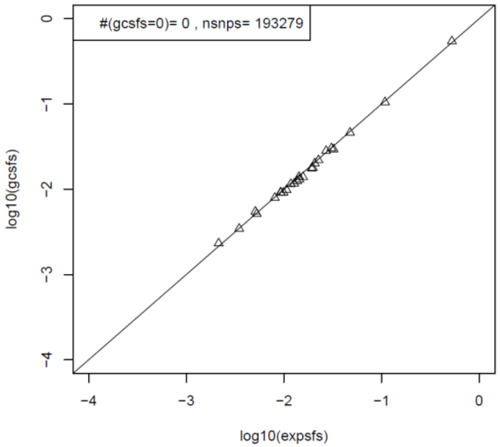
DP > 10



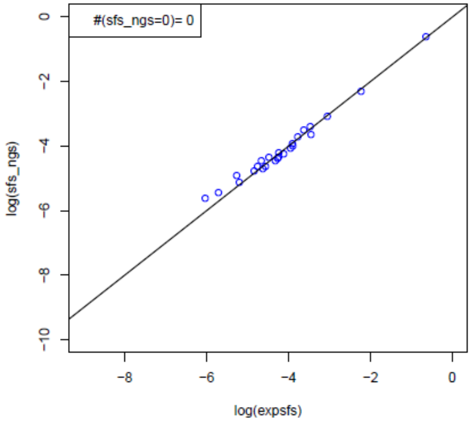
DP > 15



DP > 20



SFS accounting for genotype uncertainty (ANGSD)

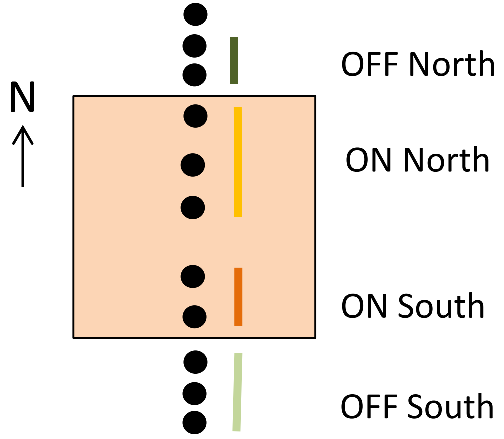


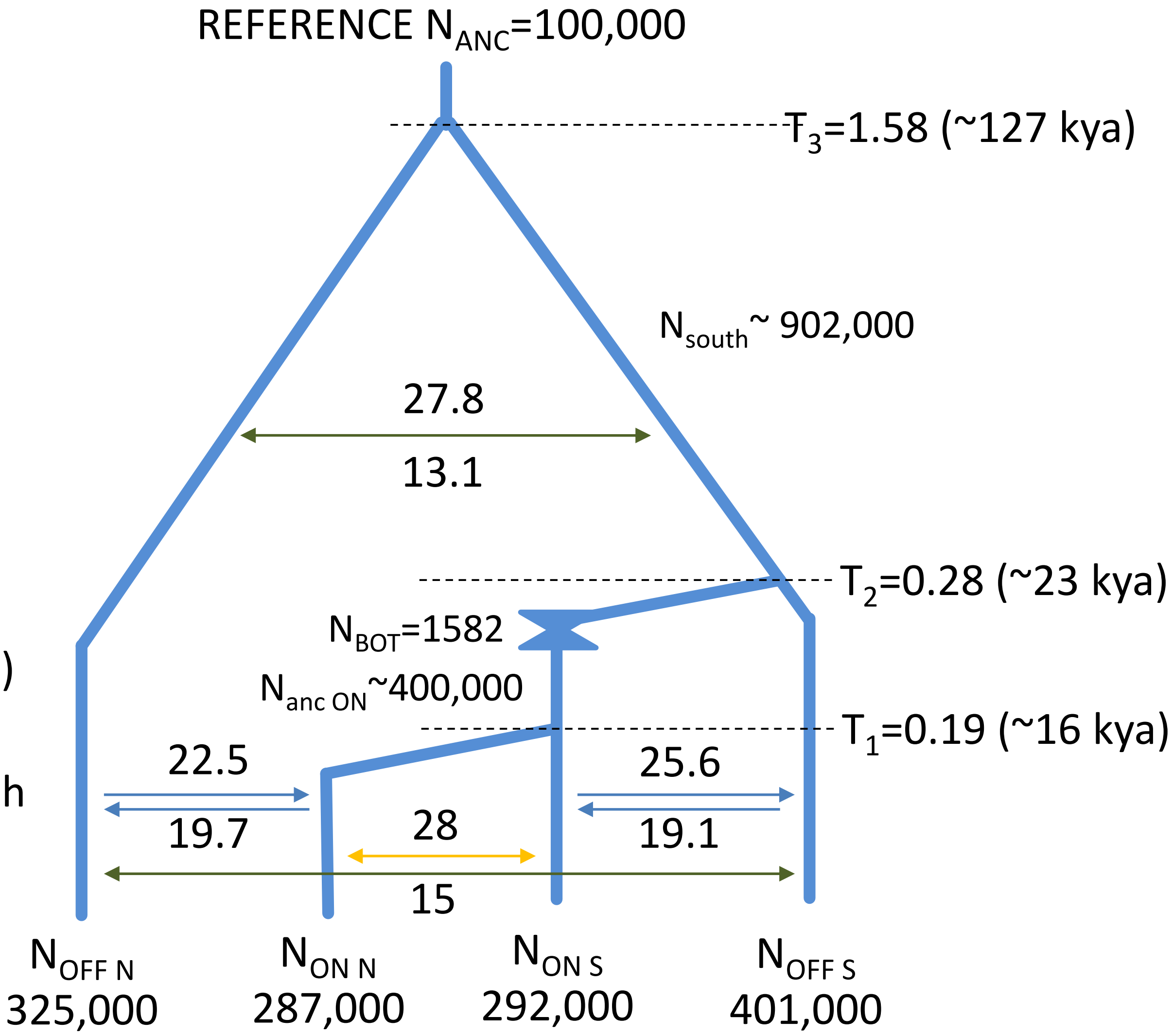
Simulated 2 pops SFS sampling 4 diploids from each pop, 200000 SNPs, mean coverage=**10x**, error rate=0.01. Simulated with correlated allele frequencies model ($F_{ST}$=(0.275, 0.01))

With DP>15 we have a very good approximation to the correct SFS, even when using the called genotypes

# Effect of HW filtering on demographic estimates

## Removing sites with HWE excess and deficit leads to different estimates

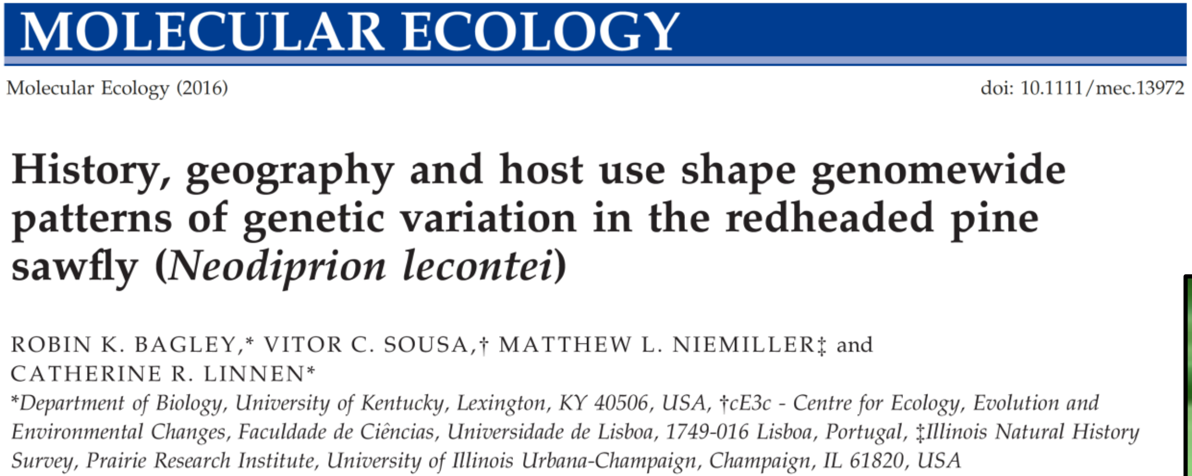




- High migration between all groups of populations (2Nm~20)
- No evidence of a strong bottleneck signal associated with colonization of SH

**DP>15 (5 diploids per group) 100,127 SNPs**

# Sawflies and RAD data



MOLECULAR ECOLOGY

Molecular Ecology (2016) doi: 10.1111/mec.13972

## History, geography and host use shape genomewide patterns of genetic variation in the redheaded pine sawfly (*Neodiprion lecontei*)

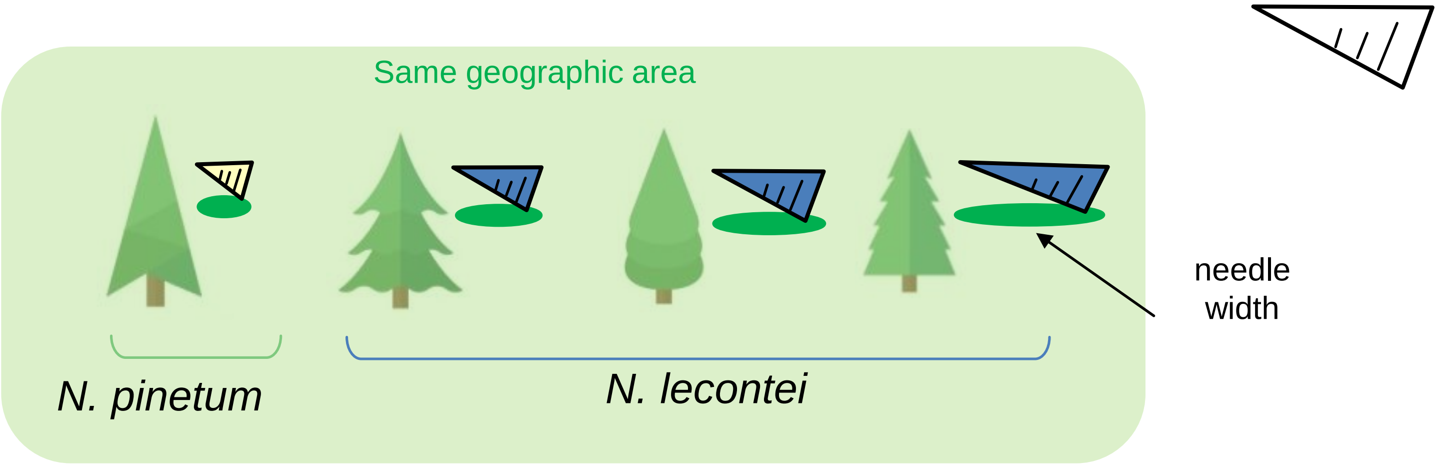
ROBIN K. BAGLEY,* VITOR C. SOUSA,† MATTHEW L. NIEMILLER‡ and CATHERINE R. LINNEN*

**Department of Biology, University of Kentucky, Lexington, KY 40506, USA, †cE3c - Centre for Ecology, Evolution and Environmental Changes, Faculdade de Ciências, Universidade de Lisboa, 1749-016 Lisboa, Portugal, ‡Illinois Natural History Survey, Prairie Research Institute, University of Illinois Urbana-Champaign, Champaign, IL 61820, USA*

# Sawflies *Neodiprion lecontei*

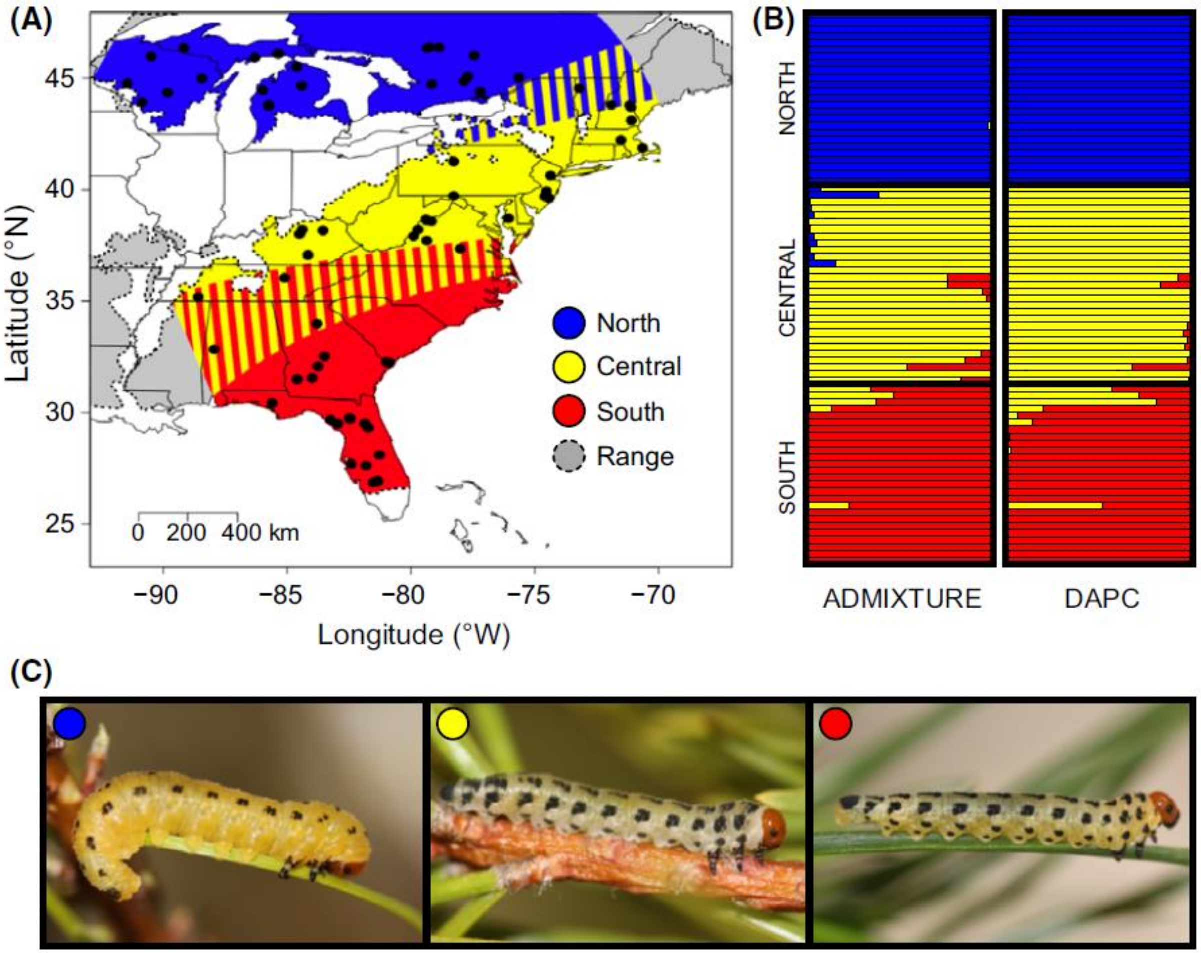
- Hymenoptera
- Plant-feeding insects
- Pine tree specialists

Ovipositor (saw)



# ddRAD seq data

- 80 individuals from 77 localities and 13 host species
- 100 bp paired-end reads, mapped to reference genome of *N. lencontei*
- Depth of coverage filter DP>10

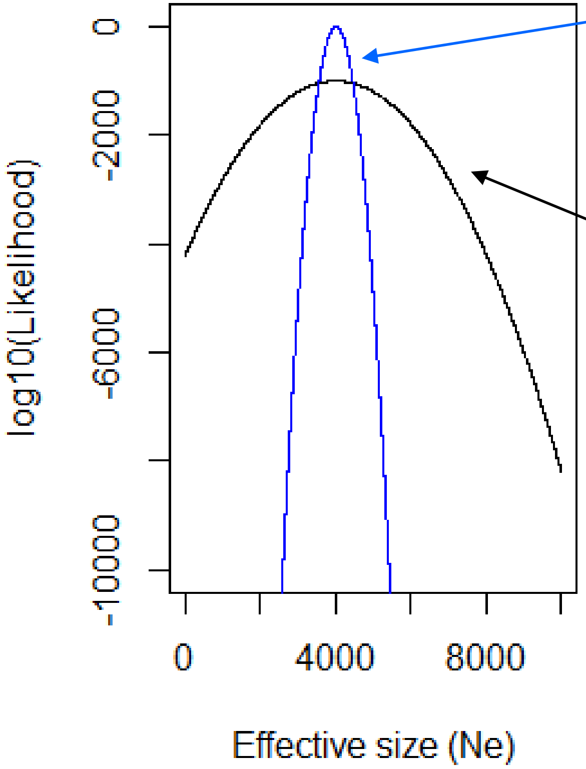


Given the detected three groups (North, Central, South):

- What is the the population tree topology?
- What are the split times?
- What are the migration levels among groups?

# Comparing models with composite likelihoods

- Fastsimcoal2 likelihood is “correct” if all SNPs are independent
- We can then compare the model likelihoods using Akaike Information Criterion (AIC)



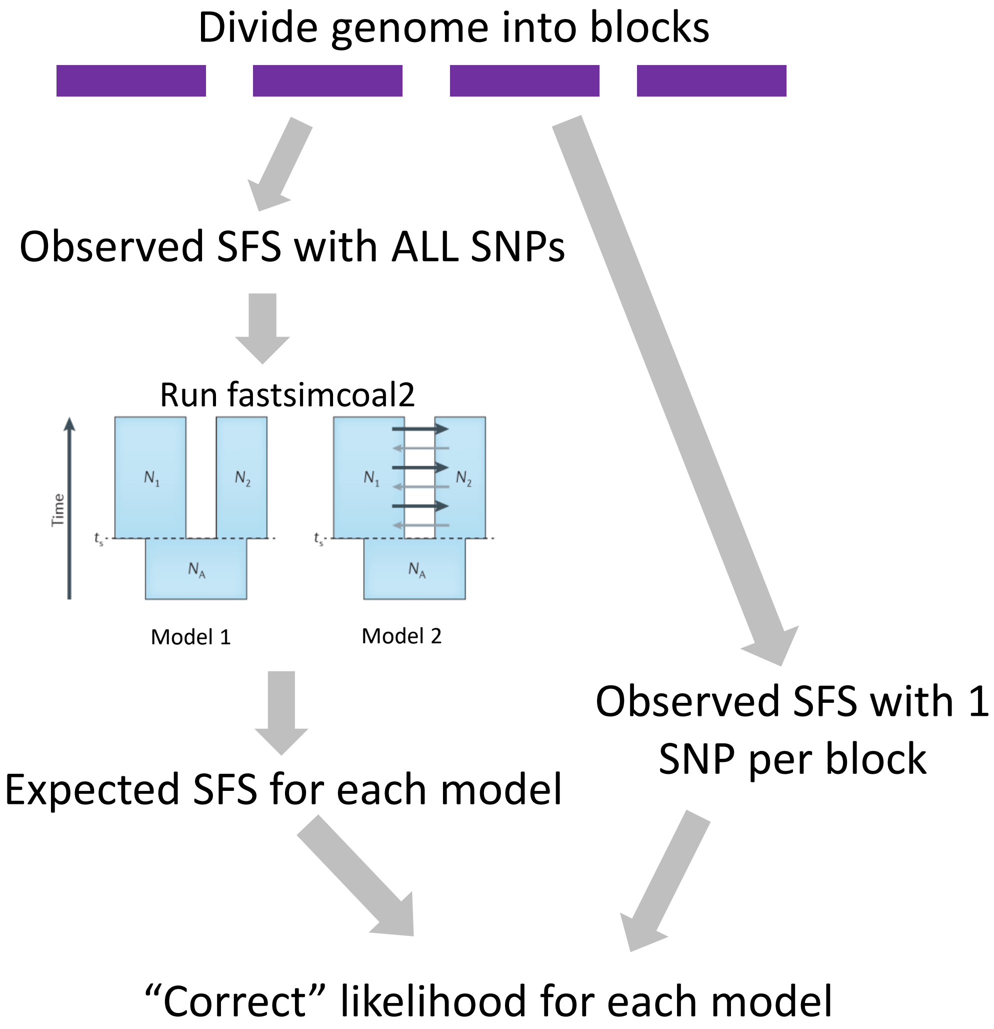
Composite likelihood (assuming linked sites are independent)

“correct” likelihood (all sites are actually independent)

Composite likelihood provide unbiased maximum likelihood parameter estimates, but the likelihoods are inflated

# A strategy to compare models

1. Divide the dataset into LD blocks.
2. Create a dataset with all SNPs (including linked SNPs)
3. For each model, obtain the parameters that maximize the likelihood (this is ok even with linked sites!) and the corresponding expected SFS
4. Create a dataset with “independent” SNPs (1 SNP per RAD tag)
5. Given the expected SFS of each model, compute the “correct” likelihood for each model with the dataset with independent SNPs
6. Compare models with AIC



Divide genome into blocks

Observed SFS with ALL SNPs

Run fastsimcoal2

Model 1

Model 2

Expected SFS for each model

Observed SFS with 1 SNP per block

“Correct” likelihood for each model

# Comparing alternative models

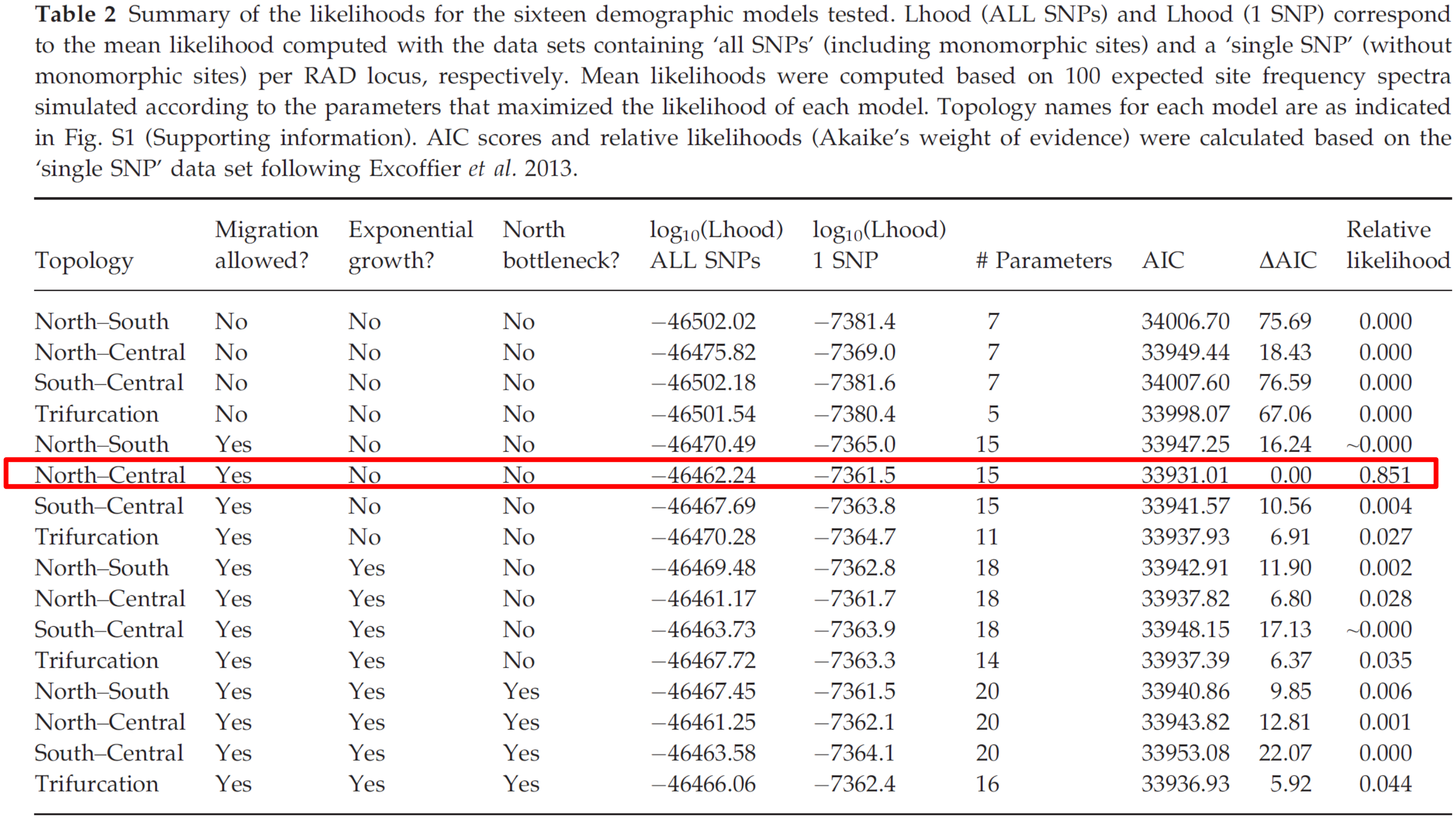
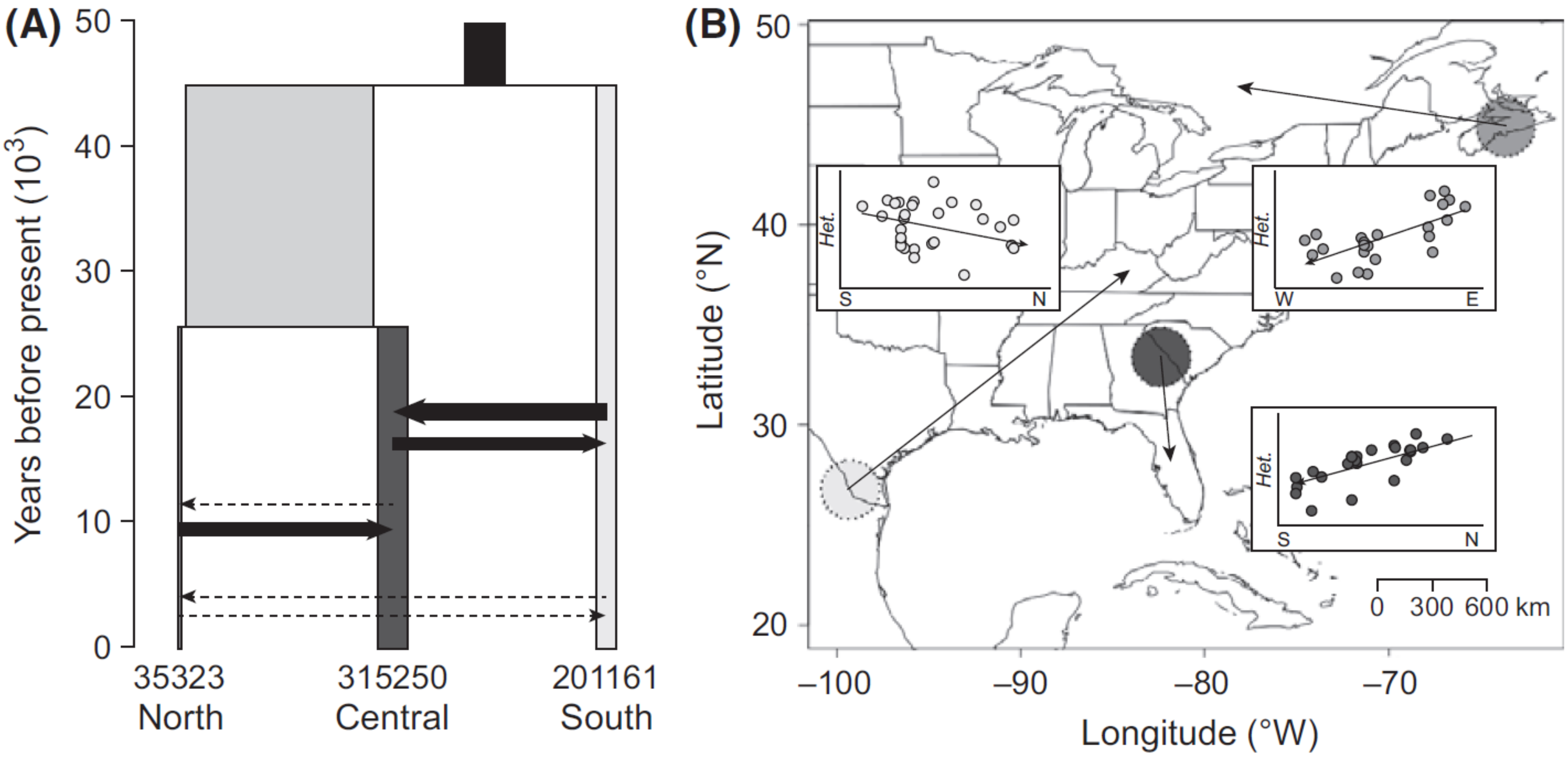


**Table 2** Summary of the likelihoods for the sixteen demographic models tested. Lhood (ALL SNPs) and Lhood (1 SNP) correspond to the mean likelihood computed with the data sets containing 'all SNPs' (including monomorphic sites) and a 'single SNP' (without monomorphic sites) per RAD locus, respectively. Mean likelihoods were computed based on 100 expected site frequency spectra simulated according to the parameters that maximized the likelihood of each model. Topology names for each model are as indicated in Fig. S1 (Supporting information). AIC scores and relative likelihoods (Akaike's weight of evidence) were calculated based on the 'single SNP' data set following Excoffier *et al*. 2013.

| Topology | Migration allowed? | Exponential growth? | North bottleneck? | $\log_{10}$(Lhood) ALL SNPs | $\log_{10}$(Lhood) 1 SNP | # Parameters | AIC | ΔAIC | Relative likelihood |
|---|---|---|---|---|---|---|---|---|---|
| North–South | No | No | No | −46502.02 | −7381.4 | 7 | 34006.70 | 75.69 | 0.000 |
| North–Central | No | No | No | −46475.82 | −7369.0 | 7 | 33949.44 | 18.43 | 0.000 |
| South–Central | No | No | No | −46502.18 | −7381.6 | 7 | 34007.60 | 76.59 | 0.000 |
| Trifurcation | No | No | No | −46501.54 | −7380.4 | 5 | 33998.07 | 67.06 | 0.000 |
| North–South | Yes | No | No | −46470.49 | −7365.0 | 15 | 33947.25 | 16.24 | ~0.000 |
| North–Central | Yes | No | No | −46462.24 | −7361.5 | 15 | 33931.01 | 0.00 | 0.851 |
| South–Central | Yes | No | No | −46467.69 | −7363.8 | 15 | 33941.57 | 10.56 | 0.004 |
| Trifurcation | Yes | No | No | −46470.28 | −7364.7 | 11 | 33937.93 | 6.91 | 0.027 |
| North–South | Yes | Yes | No | −46469.48 | −7362.8 | 18 | 33942.91 | 11.90 | 0.002 |
| North–Central | Yes | Yes | No | −46461.17 | −7361.7 | 18 | 33937.82 | 6.80 | 0.028 |
| South–Central | Yes | Yes | No | −46463.73 | −7363.9 | 18 | 33948.15 | 17.13 | ~0.000 |
| Trifurcation | Yes | Yes | No | −46467.72 | −7363.3 | 14 | 33937.39 | 6.37 | 0.035 |
| North–South | Yes | Yes | Yes | −46467.45 | −7361.5 | 20 | 33940.86 | 9.85 | 0.006 |
| North–Central | Yes | Yes | Yes | −46461.25 | −7362.1 | 20 | 33943.82 | 12.81 | 0.001 |
| South–Central | Yes | Yes | Yes | −46463.58 | −7364.1 | 20 | 33953.08 | 22.07 | 0.000 |
| Trifurcation | Yes | Yes | Yes | −46466.06 | −7362.4 | 16 | 33936.93 | 5.92 | 0.044 |

Joint 3D minor allele frequency SFS (11,617 SNPs – ALL SNPs; 4,478 SNPs – 1 SNP per RAD tag)

# Estimates favors a scenario where North and Central diverged more recently with asymmetric gene flow



The inferred population tree topology and divergence times are consistent with divergence and range expansion from different refugia after LGM

3 pairwise 2D minor allele frequency SFS (15,230 SNPs)

# Summary

- Fastsimcoal2 can be applied to RAD seq data
- We used a strategy to obtain (as close as possible) the “correct” likelihood by dividing the data into blocks, inferring the expected SFS for each model with ALL SNPs, and then re-computing the “true” likelihood with independent SNPs (1 SNP per block)
- Despite the reduced number of SNPs we were able to discriminate models based on their likelihoods

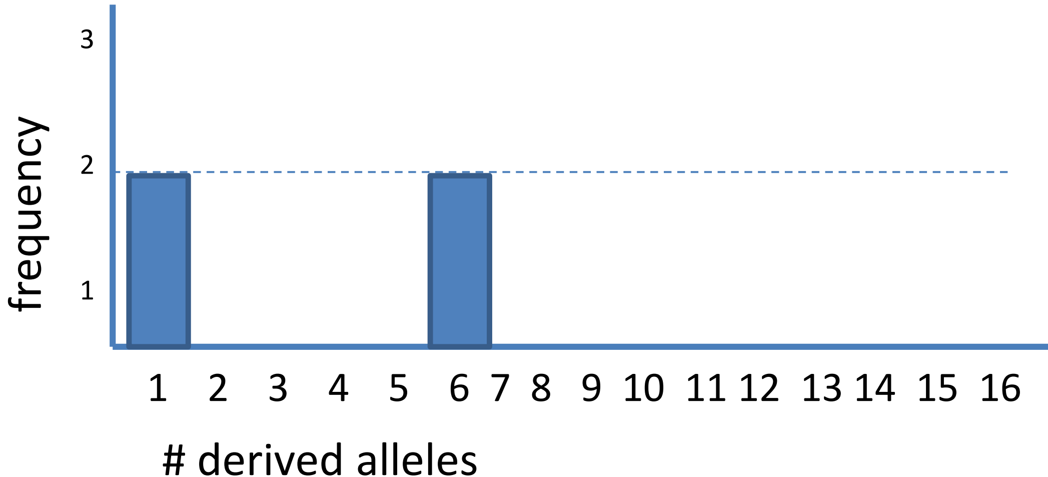
# Protocol for model comparison based on AIC when we have independent SNPs

- Get the observed SFS
- Define the alternative models
- Perform 50-100 runs under each model
- Select the runs with maximum likelihood under each model
- Compute the AIC (Akaike information critera) for each model
- Select the model with minimum AIC

# Estimating SFS from observed data

- How to deal with missing data?

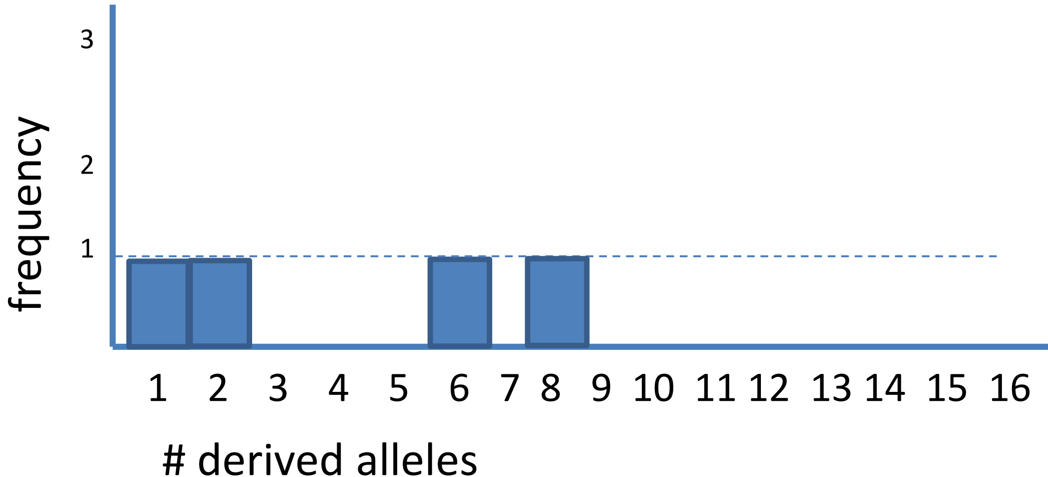
| | Freq. derived | Sample size | Rel. freq |
|---|---|---|---|
| SNP1 | 1 | 16 | 1/16 |
| SNP2 | 6 | 12 | 1/2 |
| SNP3 | 1 | 12 | 1/12 |
| SNP4 | 6 | 16 | 3/8 |



# Estimating SFS from observed data

- How to deal with missing data?

| | Freq. derived | Sample size | Rel. freq |
|---|---|---|---|
| SNP1 | 1 | 16 | 1/16 |
| SNP2 | 6 | 12 | 1/2 |
| SNP3 | 1 | 8 | 1/12 |
| SNP4 | 6 | 16 | 3/8 |

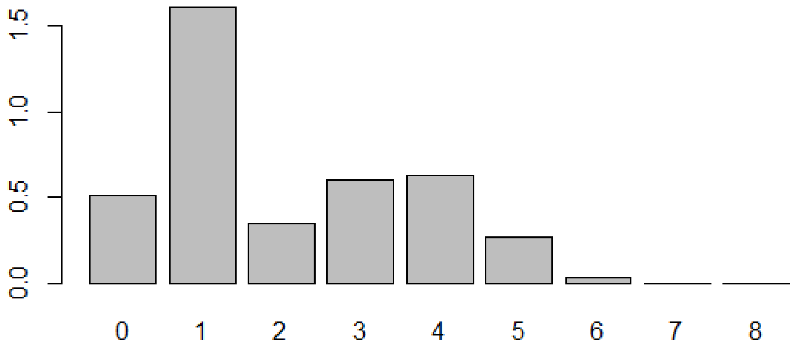


# Estimating SFS from observed data

- How to deal with missing data?

- Solution:
  - Find minimum sample size
  - Resample without replacement

| | Freq. derived | Sample size | Rel. freq |
|---|---|---|---|
| SNP1 | 1 | 16 | 1/16 |
| SNP2 | 6 | 12 | 1/2 |
| SNP3 | 1 | 8 | 1/12 |
| SNP4 | 6 | 16 | 3/8 |



Gavel et al. (2014) PNAS

# FASTSIMCOAL2 INPUT FILES

Vitor Sousa

vmsousa@fc.ul.pt

Cesky Krumlov 2018

# Examples of observed SFS

**1PopExpInst20Mb_DAFpop0.obs**

```
1 observations
d0_0      d0_1      d0_2  d0_3  d0_4  d0_5  d0_6  d0_7  d0_8  d0_9  d0_10
19973842  24630     810   173   145   111   88    84    61    56    0
```

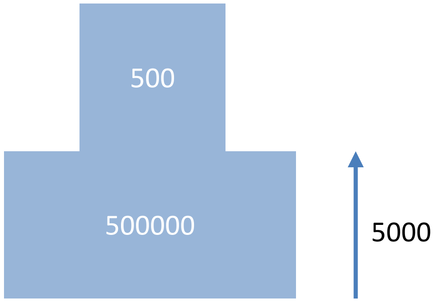
**2PopDivMigr20Mb_jointDAFpop1_0.obs**

```
1 observations
          d0_0      d0_1  d0_2  d0_3  d0_4  d0_5
d1_0      19985747  8350  1628  360   62    8
d1_1      9660      0     0     0     0
d1_2      4790      0     0     0     0
d1_3      3280      0     0     0     0
d1_4      2490      0     0     0     0
d1_5      1760      13    18    13    19    0
```

**2PopDiv20Mb_jointDAFpop1_0.obs**

```
1 observations
          d0_0      d0_1  d0_2  d0_3  d0_4  d0_5
d1_0      19985547  8211  1415  316   55    10
d1_1      1266      101   37    16    5     1
d1_2      61142     20    8     2     0
d1_3      48631     12    5     0     0
d1_4      47915     9     2     3     1
d1_5      1189      46    22    19    18    0
```

# Parameter estimation settings files



1PopExpInst20Mb

## Additional files necessary to estimate parameters

### Estimation file

**1PopExpInst20Mb/1PopExpInst20Mb.est**

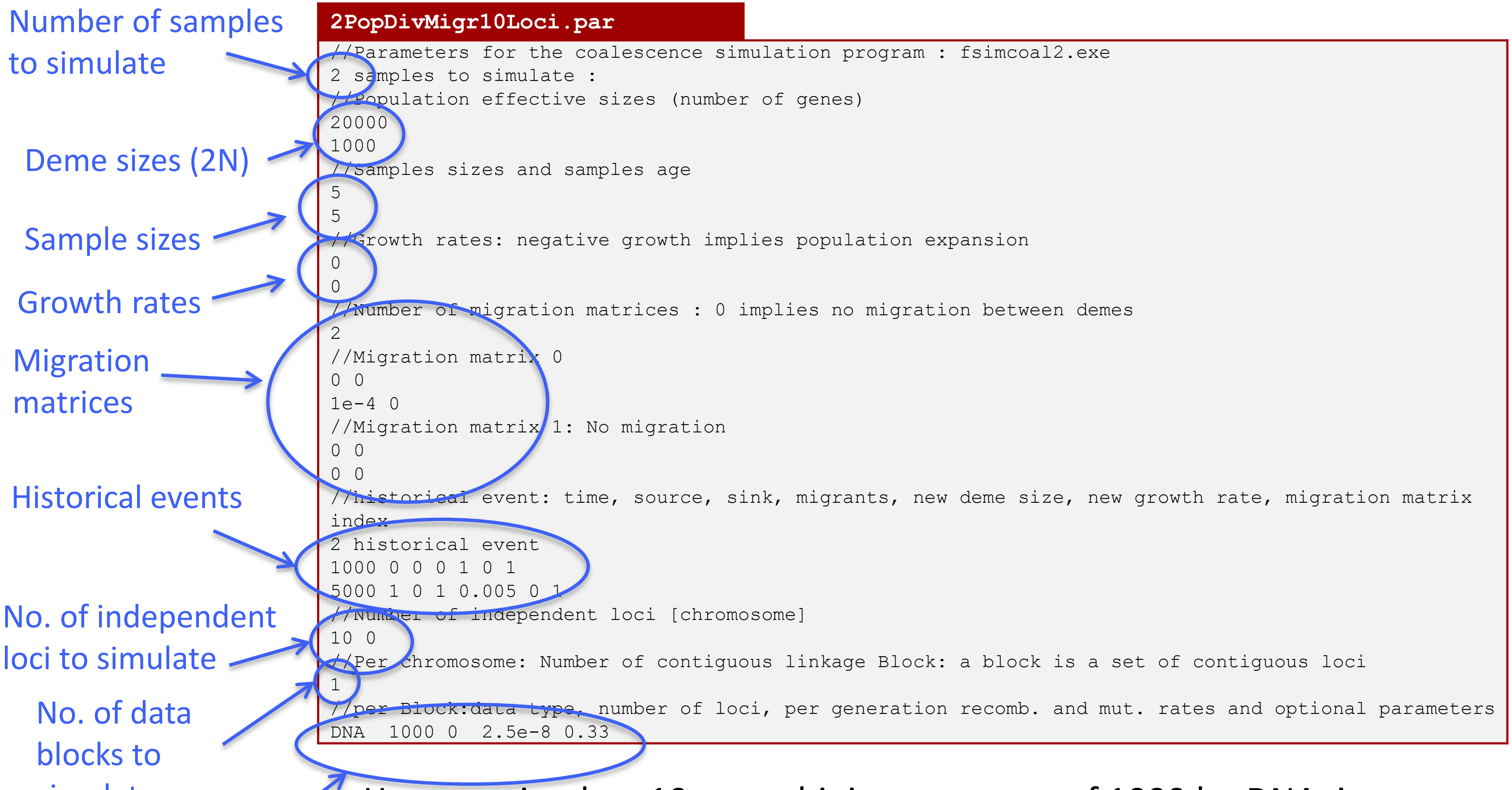
```
// Search ranges and rules file
// ****************************

[PARAMETERS]
//#isInt? #name   #dist.#min  #max
//all Ns are in number of haploid individuals
1  NPOP       logunif  1000   1e7   output
1  NANC       logunif  10     1e5   output
1  TEXP       unif     10     1e5   output

[RULES]

[COMPLEX PARAMETERS]

0  RESIZE   = NANC/NPOP      hide
```

### Template file

**1PopExpInst20Mb/1PopExpInst20Mb.tpl**

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
1 samples to simulate :
//Population effective sizes (number of genes)
NPOP
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
TEXP 0 0 0 RESIZE 0 0
//Number of independent loci [chromosome]
1 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
1
//per Block:data type, number of loci, per generation recombination and mutation rates and optional parameters
FREQ  1   0   2.5e-8 OUTEXP
```

# INPUT files for fastsimcoal2:
# Defining an evolutionary model with PAR files



**2PopDivMigr10Loci.par**

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
20000
1000
//Samples sizes and samples age
5
5
//Growth rates: negative growth implies population expansion
0
0
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0 0
1e-4 0
//Migration matrix 1: No migration
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
index
2 historical event
1000 0 0 0 1 0 1
5000 1 0 1 0.005 0 1
//Number of independent loci [chromosome]
10 0
//Per chromosome: Number of contiguous linkage Block: a block is a set of contiguous loci
1
//per Block:data type, number of loci, per generation recomb. and mut. rates and optional parameters
DNA  1000 0  2.5e-8 0.33
```

Number of samples to simulate

Deme sizes (2N)

Sample sizes

Growth rates

Migration matrices

Historical events

No. of independent loci to simulate

No. of data blocks to simulate

Definition of genetic data type to simulate

Here we simulate 10 recombining segments of 1000 bp DNA, in two populations of sizes 20000 and 1000 having diverged 5000 generations ago from a small population of size 100

# TPL files

TPL are like PAR files, but the actual parameter values are replaced by parameter tags.

These files are very important! Check carefully all the definitions. Errors in the TPL file are difficult to detect and imply the model specification is incorrect! This means that all inferences will be wrong, and also that all parameter estimates will be incorrect!

**Defining population sizes and sample sizes**

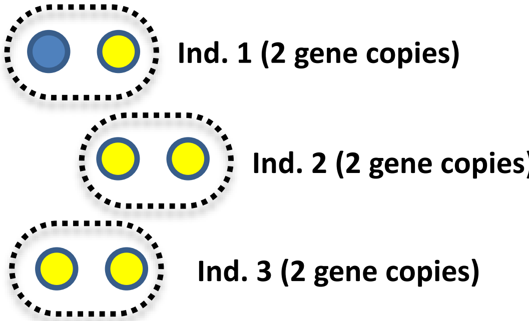
**2PopDivMigr10Loci.par**

```
//Parameters for the coalescence simulation program : fsimcoal2.exe
2 samples to simulate :
//Population effective sizes (number of genes)
NPOP1
NPOP2
//Samples sizes and samples age
6
6
//Growth rates: negative growth implies population expansion
0
0
```

Parameter tags

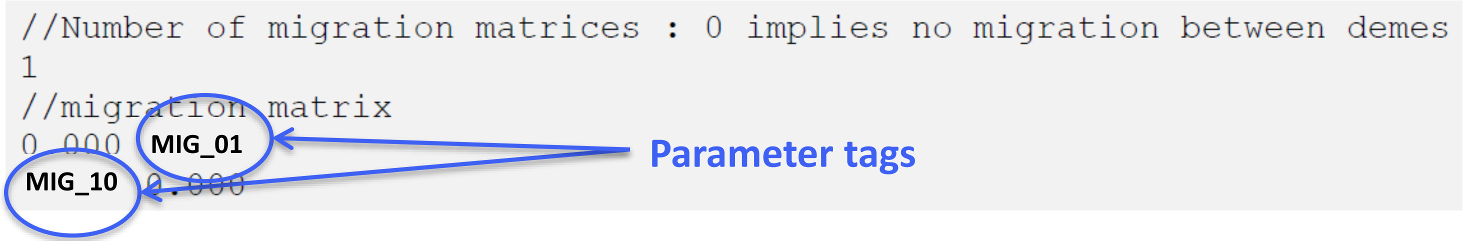
Population effective sizes are given in number of gene copies. For a diploid species with N=500 individuals, this corresponds to a 2N=1000 gene copies, as each individual carries two gene copies at any given site.

The sample size is also given in gene copies. The value of 6 means that we sampled 3 diploid individuals.



# TPL files

## MIGRATION



```
//Number of migration matrices : 0 implies no migration between demes
1
//migration matrix
0.000 MIG_01
MIG_10 0.000
```

The migration matrix can be asymmetric, and in the case the entry $m_{ij}$ list the **migration rates backward in time** from population *i* to population *j*. The above-mentioned matrix states that, for each generation backward in time, any gene from population 0 has probability MIG_01 to be sent to population 1, and that a gene from population 1 has a probability MIG_10 to move to population 0.

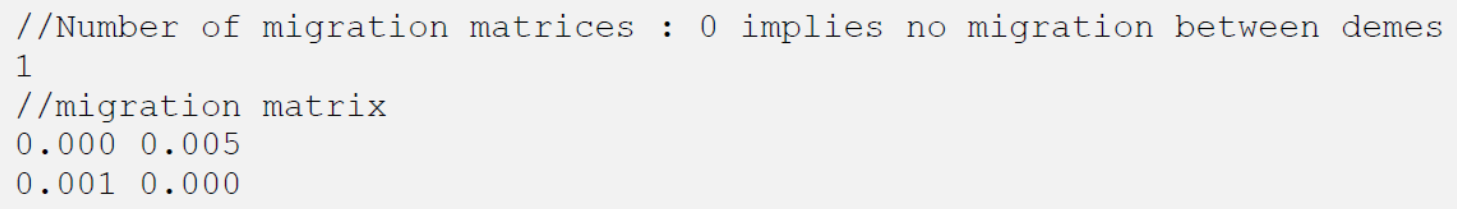
If no migration matrix is defined, no migration is assumed between populations.

**1PopStationary10Loci.par**

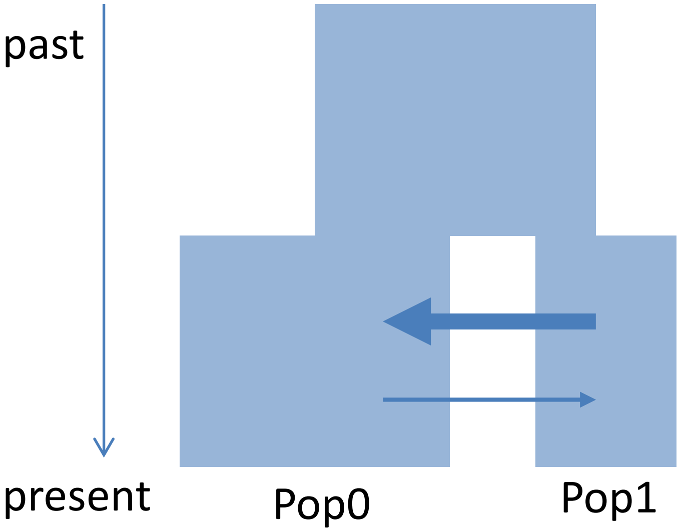
```
//Number of migration matrices : 0 implies no migration between demes
0
```

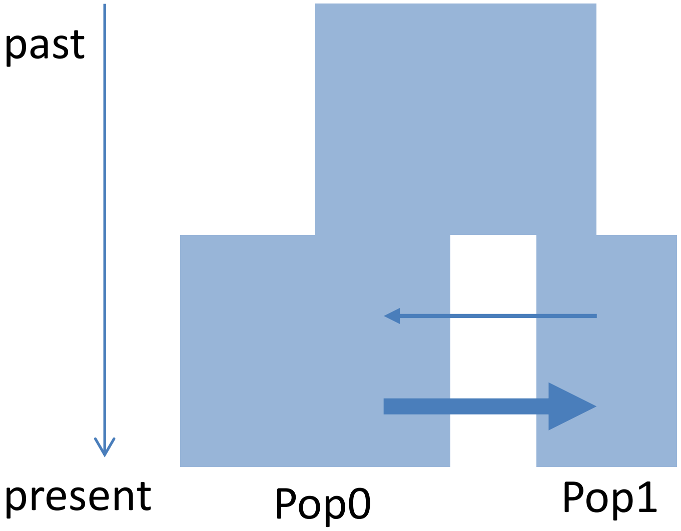
# A note on looking backward in time

Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?



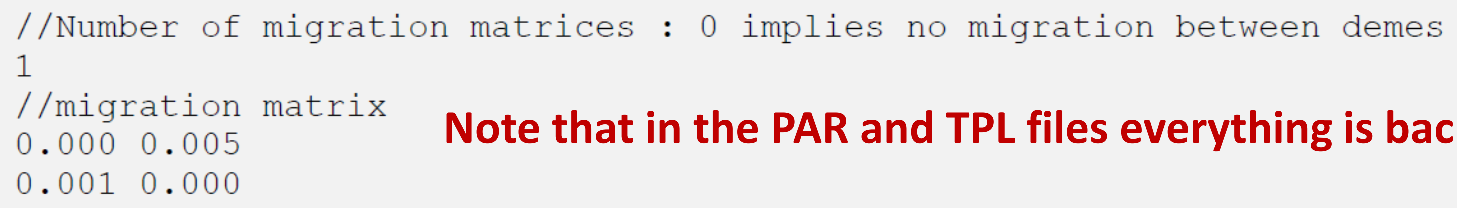
```
//Number of migration matrices : 0 implies no migration between demes
1
//migration matrix
0.000 0.005
0.001 0.000
```





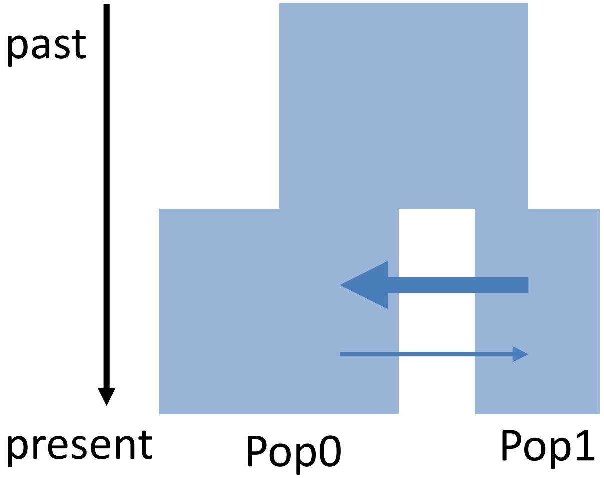
# A note on looking backward in time

Assuming that we look forward in time and that the size of the arrows are proportion to the migration rate, to what model does the following migration matrix corresponds to?

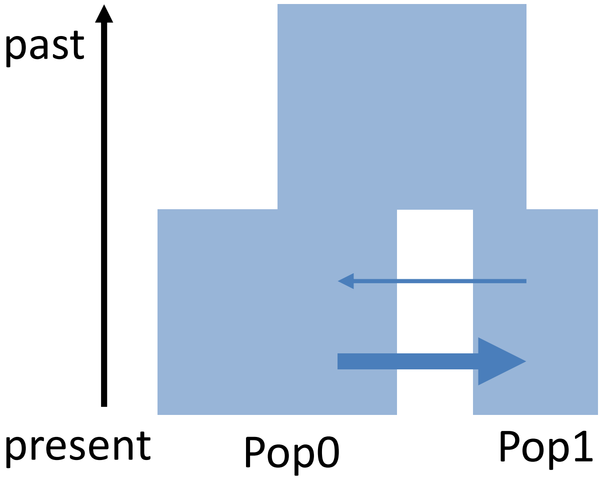


```
//Number of migration matrices : 0 implies no migration between demes
1
//migration matrix
0.000 0.005
0.001 0.000
```

**Note that in the PAR and TPL files everything is backward in time!!**



**This is the correct model forward in time, meaning there are more migrants moving from pop1 to pop0 each generation.**



**Backward in time this is the model. Lineages are more likely to move from pop0 to pop1.**

# Historical events in fastsimcoal2

Historical events can be used to:

- Change the size of a given population
- Change the growth rate of a given population
- Change the migration matrix to be used between populations
- Move a fraction of the genes of a given population to another population. This amounts to implementing a (stochastic) admixture or introgression event.
- Move all genes from a population to another population. This amounts to fusing two populations into one looking backward in time.
- One or more of these events at the same time

Defining the historical events is crucial to have a correct model!

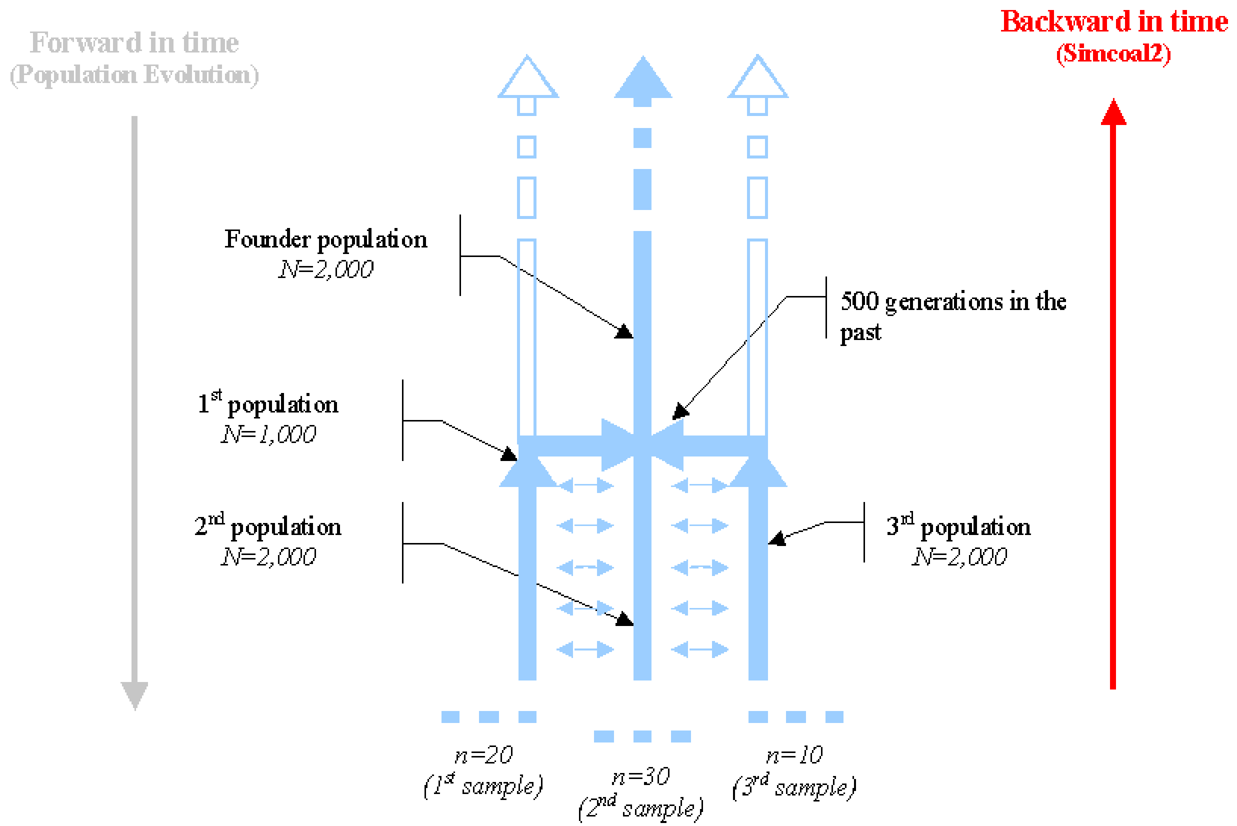
# Historical events (backward in time)

Each historical event is coded with a line with the following arguments

**time, source, sink, migrants, new deme size, new growth rate, migration matrix index**

| | | | | | | |
|---|---|---|---|---|---|---|
| **500** | **0** | **1** | **1** | **1** | **0** | **1** |
| **500** | **2** | **1** | **1** | **1** | **0** | **1** |

500 generations ago, 100% (**migrants=1.0**) of lineages in **pop0** (**source =0**) migrated to **pop1** (**sink=1**). The size of the sink (pop1) remained the same (**new deme size=1.0**, i.e. N2=2000). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.



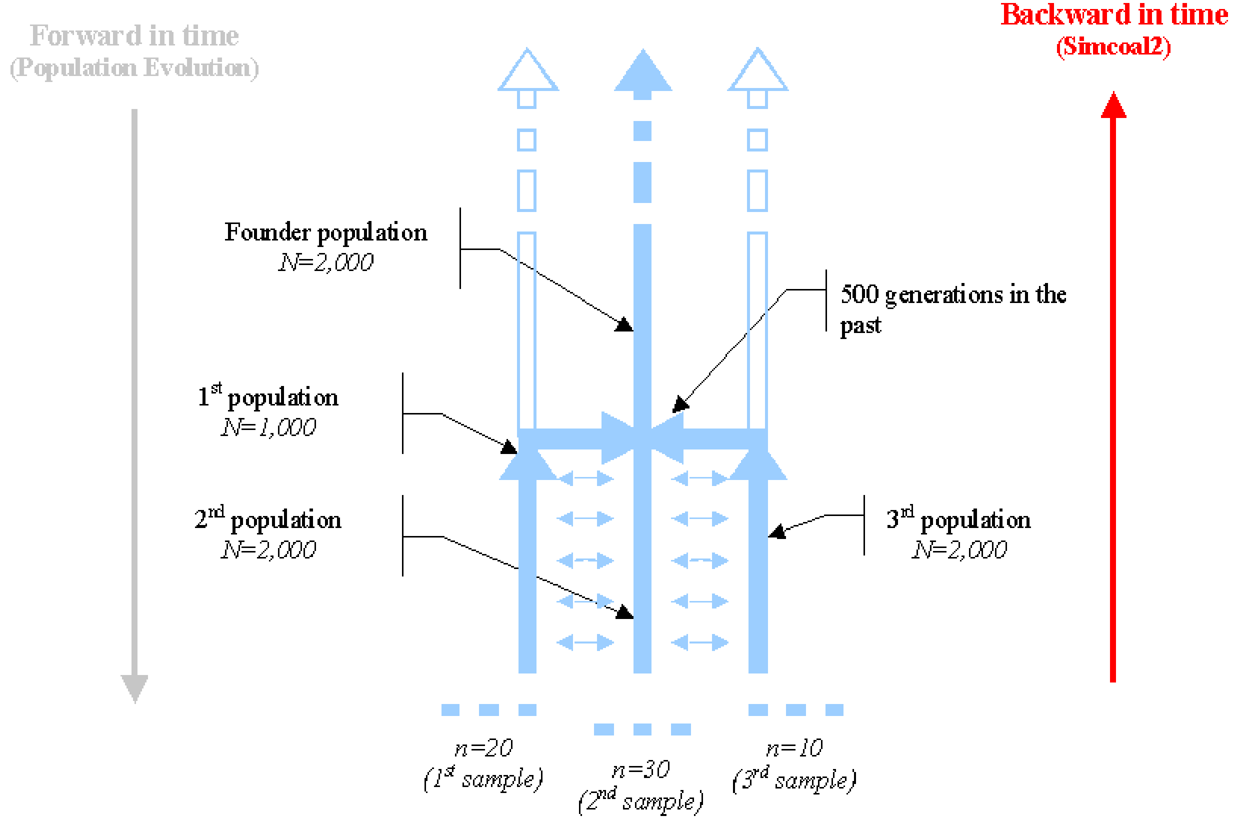
# Historical events (backward in time)

Each historical event is coded with a line with the following arguments

**time, source, sink, migrants, new deme size, new growth rate, migration matrix index**

**500 0 1 1 1 0 1**

**500 2 1 1 1 0 1**



**500** generations ago, **100%** of lineages (**migrants=1.0**) in **pop2** (**source =2**) migrated to **pop1** (**sink=1**). The size of the sink (pop1) remained the same (**new deme size=1.0**, i.e. N2=2000). The new growth rate is zero. The migration rate that is active after the event is given in the migration matrix 1.

# Historical events in fastsimcoal2

## Change the size of a given population

**1PopContrInst10Loci.par**

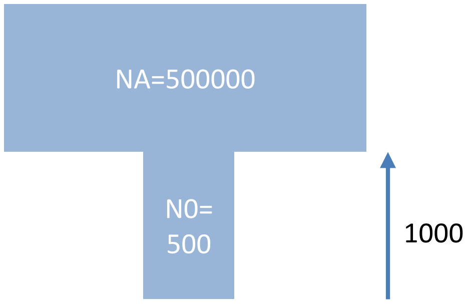
```
//Parameters for the coalescence simulation program : fsimcoal2.exe
1 samples to simulate :
//Population effective sizes (number of genes)
1000
//Samples sizes and samples age
10
//Growth rates: negative growth implies population expansion
0
//Number of migration matrices : 0 implies no migration between demes
0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix index
1 historical event
1000 0 0 0 1000 0 0
```

- 1000 generations ago, 0% (migrants=0) of lineages in pop0 (source) migrated to pop1 (sink). This means that 100% of lineages remained in pop0.

- The sink population (pop0) has a size 1000 larger after the event (new size=1000). Given that N0=500 diploids at time zero, it implies that NA=500000 diploids.

- The migration matrix valid after the event is the migration rate 0. Since it is not defined it imples no migration.

Recent instantaneous demographic contraction



1PopContrInst10loci.par

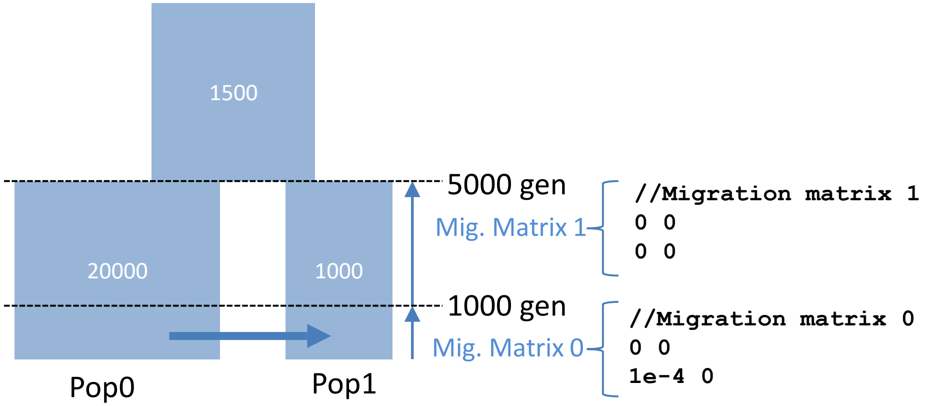
# Historical events in fastsimcoal2

## Change the migration matrix to be used between populations

**2PopDivMigr10Loci.par**

```
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0  0
1e-4  0
//Migration matrix 1: No migration
0  0
0  0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
index
2 historical event
1000  0  0  0  1  0  1
5000  1  0  1  1.5  0  1
```

- At generation 1000 in the past, 0% (migrants=0) of lineages migrated from pop0 (source=0) to pop1 (sink=0).
- After the historical event, the deme size of the sink population (pop1) remained the same (new deme size=1).
- After the historical event the growth rate was set to zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.



```
//Migration matrix 1
0  0
0  0
```

```
//Migration matrix 0
0  0
1e-4  0
```

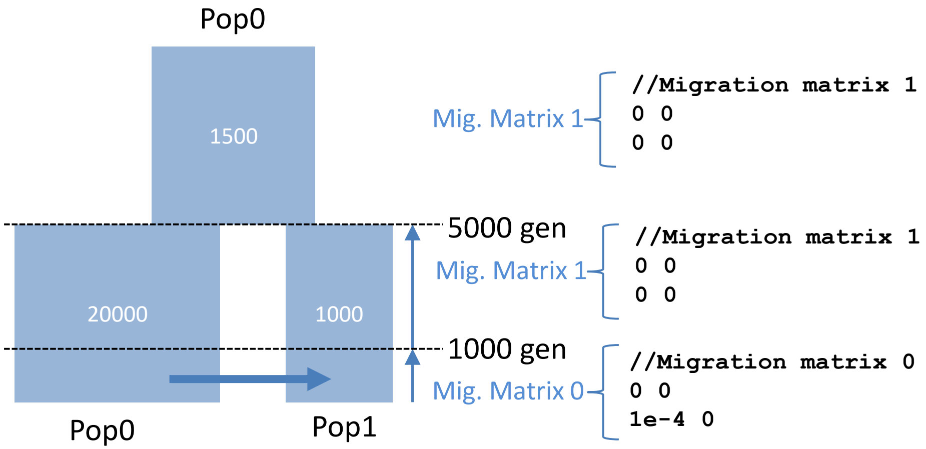
# Historical events in fastsimcoal2

## Population split (merge populations going backwards in time)

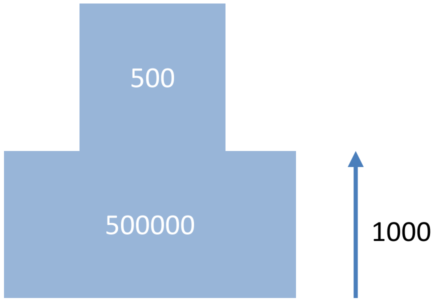
**2PopDivMigr10Loci.par**

```
//Number of migration matrices : 0 implies no migration between demes
2
//Migration matrix 0
0 0
1e-4 0
//Migration matrix 1: No migration
0 0
0 0
//historical event: time, source, sink, migrants, new deme size, new growth rate, migration matrix
index
2 historical event
1000 0 0 0 1 0 1
5000 1 0 1 0.075 0 1
```

- At generation 5000 in the past, 100% (migrants=1) of lineages migrated from pop1 (source=1) to pop0 (sink=0).
- After the population split, the deme size of the sink population (pop0) is 1500 (new deme size=1500/20000=0.075).
- After the historical event the growth rate of the sink population pop0 is zero.
- After the historical event the migration rate matrix was set to matrix 1, i.e. no migration between populations.



# Launching parameter estimations



Command line to estimate parameters

```
fsc22.exe -t 1PopExpInst20Mb.tpl -e 1PopExpInst20Mb.est
          -M0.001 -d -n100000 -N100000 -l5 -L20 -q -c4
```

Template file name

Estimation file name

No. of threads

Do ML estimation
And stop criterion

Use derived SFS

Minimum (-n) and maximum (-N) number of simulations to estimate the expected SFS

Minimum (-l) and maximum (-L) number of iterations to perform

Observed SFS file must have the same name as template file and extension _DAFpop0.obs. e.g. `1PopExpInst20Mb_DAFpop0.obs`